

In praise of immigration

The United States is a nation of immigrants — and nowhere more so than in the lab. Yet officials of the federal government don't seem to recognize that the country's scientific strength depends in large part on foreign talent.

After the atrocities of 11 September 2001, it was inevitable that the US government would tighten up its procedures for letting foreign nationals into the country. For the Department of State and consulates abroad, security is now the watchword.

No one disputes the need to exclude potential terrorists, but the resulting controls on immigration have become an unwieldy mess, arbitrarily ensnaring individuals who would previously have been welcomed into the country with open arms. Visiting scientists, perhaps more than any other group, have experienced these policies first-hand — leading to missed conferences, lost lab time and delayed graduations for large numbers of scientists and students.

Some of the affected individuals have been through a nightmarish experience, their careers thrown seriously off track. But as *Nature's* investigation into the issue reveals, the greatest damage may ultimately be suffered by the US scientific enterprise (see page 190). If the world's brightest young scientists turn instead to other academic destinations, the quality of research in US labs will suffer.

You might expect to find science-oriented officials within the federal government speaking up to stress the important contribution made by foreign-born scientists, many of whom take up permanent residence in the United States. But in the most part, they have failed to do so. Even Elias Zerhouni, the Algerian-born director of the National Institutes of Health, has had little to say in public on the issue.

Some comments by senior officials have added to the impression that foreign scientists aren't valued. At a press conference in November on the status of the US scientific workforce, for instance, Rita Colwell, director of the National Science Foundation, told reporters: "We must end our addictive dependence on foreign workers." Her words were intended to bolster support for US science education, but they expose a widely held view that researchers from abroad are a stopgap that should be replaced by home-grown talent.

History teaches us a different lesson. The first scientific Nobel

prize won by the United States went in 1907 to Albert Michelson, a Prussian-born physicist whose measurement of the velocity of light inspired Einstein's theory of relativity. Since then, immigrant scientists have accounted for more than a quarter of the United States' Nobels in physics, chemistry and medicine. These scientists' journeys from their homelands to the United States were not simply about securing superior funding and laboratory equipment. Many fled discrimination, war and genocide. Others were politically active in countries where dissenters were imprisoned or executed. All saw the United States as a land of freedom and opportunity.

The experiences of foreign scientists currently attempting to enter the United States are eroding that perception. Visiting scientists are increasingly finding themselves in situations that are stressful or humiliating. Even more worrying are reports of hate crimes committed against Iranian and Arab students.

In contrast, over the past decade, other Western nations have worked to make their societies more receptive to talented foreigners by loosening immigration laws for technical workers and trying to combat prejudice in their societies. In the twenty-first century, the United States is just one of many destinations.

The US scientific leaders who recognize the important contribution made by foreign scientists are split into two camps. The optimists point out that about 40% of the world's research dollars are spent inside US laboratories. A few visa delays will do little to change that scientific hegemony, they argue. But the pessimists fear that the world's rising scientific stars are already starting to turn their backs on the United States. Even a superpower can't afford to be complacent, they warn.

Whoever is right, federal government officials must not seem indifferent to the plight of foreign scientists. They need to take measures to ensure that these valuable assets are treated with dignity and made to feel welcome. ■

Don't fear the Robot Scientist

Contrary to first impressions, an automated system that designs its own experiments will benefit young molecular geneticists.

At first glance, it seems to render obsolete the armies of postgrads and postdocs employed in the world's molecular-genetics laboratories. In this week's issue (see page 247), a British team unveils an automated system that "originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using a laboratory robot, interprets the results to falsify hypotheses inconsistent with the data, and then repeats the cycle".

What's more, when set loose on experiments to investigate the genetic control of important metabolic pathways in yeast, it performs more cost effectively than scientifically educated human volunteers. The Robot Scientist seems to promise a future of successfully completed research projects, untouched by human mind.

Neo-luddites must be unsure whether to curse or celebrate. On one hand, they are obliged to condemn another technology that

seems to threaten established patterns of employment. But they may also be glad to see the scientific and technological elite seemingly hoist with its own petard.

The truth is rather different. The Robot Scientist does represent an important step forward, but does not spell the end for its human counterpart. The deductive steps required to design experiments for functional-genomic analyses are particularly amenable to solution by computer algorithms. And this is a field in which the deluge of data requiring explanation exceeds researchers' capacity to cope.

The team behind the Robot Scientist argues that such automation "frees scientists to make the high-level creative leaps at which they excel". Therein lies the challenge. Some lab heads still treat postgrads and postdocs as a cheap source of menial labour, rather than educating them to become tomorrow's creative research leaders. We can only hope that the Robot Scientist helps to change such attitudes. ■





Double life
National Science
Foundation's security
role raises eyebrows
p184



History lesson
Blood samples offer
clues to earlier
SARS-like infections
p185



Get connected
UK bids to boost
multidisciplinary
research
p187



Flying start
Scientists prepare to
comb the honeybee
genome
p188

Robotic missions set to benefit as US takes aim at the Moon

Tony Reichhardt, Washington

A US push to send astronauts back to the Moon would require a series of robotic precursor missions that could reinvigorate lunar science, say researchers working in the field.

The human-spaceflight plan, which was due to be announced by President Bush this week, won't be driven primarily by lunar science — its main goal will be to prepare the way for deep-space exploration. "Going back to the Moon can't be just science-driven. There have to be broader goals than that," says Jeffrey Taylor, a planetary scientist at the University of Hawaii in Honolulu.

Nonetheless, some branches of space science could benefit greatly if the effort takes off. In particular, the initiative could lend impetus to unmanned lunar-exploration projects already on the drawing board.

Taylor and Paul Spudis, a planetary scientist at Johns Hopkins University's Applied Physics Laboratory in Baltimore, for example, are collaborating on a proposal for a robotic mission to collect samples from the Moon's south pole and return them to Earth by 2010. Dubbed 'Farside', their concept will compete with other mission proposals submitted to NASA's New Frontiers programme next month. If selected, the mission would cost up to \$700 million. The best reason for sending astronauts to the Moon is to "learn to live off the planet", says Spudis. Proponents say that Farside would help architects of the manned mission to meet that goal.

Another team, led by lunar researcher Michael Duke of the Colorado School of Mines and involving NASA's Jet Propulsion Laboratory at Pasadena, California, will submit a competing proposal for the lunar sample mission called 'Moonrise'.

Before astronauts venture onto the Moon's surface, Duke envisages that another robotic lander would be needed to prospect for ice near the poles and identify regions of permanent shadow and sunlight. Previous lunar orbiters have detected hydrogen in the Moon's soil, but its form is not known. If it is held in ice it could be used to produce drinking water and fuel for lunar explorers, and converting it would be a research topic at a future lunar base, Spudis predicts.



Return trip: US plans to send astronauts back to the Moon have met with a mixed response.

Other possible scientific investigations by robots or astronauts include systematic studies of small craters that might reveal the history of impacts from meteorites and other space debris, which may have a bearing on past mass extinctions on Earth. Astronaut geologists could also collect meteorites blown onto the Moon's surface from Earth and other planets in the distant past. By one estimate, there could be as much as 200 kilograms of terrestrial material on each square kilometre of the Moon's surface.

Whether a new manned space programme will now unfold in the United States remains an open question. Editorial writers and some Democratic presidential candidates immediately attacked the project as fiscally irresponsible when an outline of it was leaked last week. Others, including Sherwood Boehlert (Republican, New York), chairman of the House Committee on Science, offered cautious support.

Bush cabinet members appeared on television over the weekend to stress the plan's frugality. It is expected to involve an increase of about \$800 million, or 5%, in NASA's annual budget for several years, but most of its

funding would come from the rapid phasing-out of the space shuttle, which costs NASA \$4 billion each year to fly. Another way to make the plan more affordable would be to rely on commercial Delta and Atlas rockets that have either already flown or are in development.

Reaction from scientists has also been mixed. While some worry that the programme could squeeze NASA's budget for areas such as astronomy, others point out that space-science spending rose sharply while the International Space Station was being built. Funding of space science and human exploration "is not a zero-sum game", says Wesley Huntress, a planetary scientist at the Carnegie Institution of Washington and former head of space science at NASA.

One of many questions facing the proposed programme is how the space station will fit in with it. Huntress and others told Congress in October that the station has little role to play in future human voyages to the Moon and Mars. But Spudis says that it might be smart politics for NASA to make use of the still-unfinished laboratory, perhaps even using it as a staging post on the way to the Moon.

NASA

Plans resurrected to raise Venice above the encroaching sea

Alison Abbott, Munich

Venice could be saved from sinking into the sea by using oil-industry technology to pump fluid underneath the city, says a team of geomechanical engineers.

Led by Giuseppe Gambolati of the University of Padua, the group published its proposal in a paper last month (A. Comerlati *et al.* *Eos* 84, 546, 552–553; 2003). The idea revisits a proposal first mooted in the 1970s, but Gambolati's team claims that advances in technology would now allow Venice to be raised safely.

The proposal will next be considered by CORILA, the consortium charged with coordinating research into the city's lagoon system.

Venice floods regularly, owing to climate-induced increases in sea levels in the lagoon and to over-extraction of groundwater, which has led to subsidence.

The Consorzio Venezia Nuova, the public authority responsible for protecting Venice, has taken various steps. These include raising the pavement in very low-lying areas, and building a controversial, huge mobile flood barrier called MOSE, whose gates will close during extreme sea surges (see *Nature* 424, 608–609; 2003). MOSE, costed at €3 billion (US\$3.8 billion), is due to come into operation in 2011 — but rising sea levels could render it ineffective well within 100 years.

Gambolati says that his plan could lift Venice by up to 30 cm in 10 years, "which would help to counter the effect of rising sea levels and extend MOSE's useful life".

His proposal involves injecting either carbon dioxide from local power stations or — more simply and cheaply — sea water, into a sandy layer 600–800 metres below the lagoon. The layer is sandwiched between clay below and 25 metres of relatively impermeable cap rock above.

Pierpaolo Campostrini, director of CORILA, says of the earlier proposals to pump fluid below Venice: "The ideas then involved pumping only 40 or 50 metres below the surface, which would have led to uneven raising."

Gambolati says that slow pumping over a period of 10 years at greater depth should allow lifting to be even. Sceptics, such as Rafael Bras, a hydrologist from the Massachusetts Institute of Technology, warn that much more information is needed to establish the idea's feasibility.

CORILA now plans a feasibility study, including geological and geophysical analyses of the sea bed in the lagoon and a test drilling site a safe distance away. ■

Intelligence law draws fire over NSF security project

Geoff Brumfiel, Washington

The US Congress has told the National Science Foundation (NSF) to coordinate research on intelligence matters. But critics say that if the research agency acts as instructed, it will violate its tradition of openly disseminated science.

The Intelligence Authorization Act for 2004, signed into law on 13 December, calls for the NSF to hold two workshops to coordinate research in the behavioural, psychological and physiological sciences for the purpose of government security evaluations. NSF officials are also required to head a committee of experts in defence, law enforcement and intelligence advising the government on research into security screening procedures, such as lie-detector tests.

The act specifies that the workshops and panel are exempt from government "sunshine laws", which would open the meetings and documents to public examination.

"This is at odds with an NSF-declared commitment to openness," protests Steven Aftergood, who directs the Project on Government Secrecy at the Federation of American Scientists in Washington. Such exemptions are increasingly routine in new legislation related to security, he says.

The NSF and the White House Office of Science and Technology Policy would receive US\$500,000 from the director of the Central Intelligence Agency to carry out the workshops and committee work. Because the final budget for intelligence is classified, officials were unable to say whether this expenditure is going ahead.

The NSF already works in areas related to



Close encounter: FBI trainees expect to work in secrecy, but scientists are taking a different view.

homeland security. Its Approaches to Combat Terrorism initiative, for example, has spent roughly US\$3.5 million on research for the intelligence community. But all this work is openly published.

The Senate intelligence committee is said to have inserted language mandating the workshops following a study by the National Academy of Sciences last year. The study had called for a research programme on security screening that would "operate under the normal rules of scientific freedom and openness to the extent possible while protecting national security". A committee spokesman was not available for comment.

NSF officials say the programme will not violate the agency's commitment to openness. "No one at the NSF imagines that this is calling for some kind of closed or classified workshop," says Curt Supplee, who directs the NSF office of legislative and public affairs. ■

Europe warned against research council

Jim Giles, London

The Royal Society has poured cold water on plans to create a central funding body for science in Europe.

In a working paper due to be released on 15 January, the influential society warns that the proposed European Research Council (ERC) might become excessively bureaucratic and prone to political influence. It also charges that the ERC's funding could ultimately come at the expense of existing, national research agencies, and that the body is unlikely to help raise Europe's research spending to US levels.

Last October, a report from science-policy experts recommended creating the ERC within three years, with an annual budget of €2

billion (US\$2.6 billion). The report said the council should be politically answerable to both the European Parliament and European Union (EU) member states, although its distribution of grants would be based purely on scientific merit (see *Nature* 425, 440; 2003).

But the Royal Society says the EU Council of Ministers, which is now considering the report, should think hard before legislating for an ERC. Julia Higgins, a chemist at Imperial College London, and a vice-president of the society, says that despite ERC supporters' talk of "new money" for the council, funds will all come from member states in the end.

"If the EU raises money from member states, that could ultimately affect their own research budgets," says Higgins. ■



SARS carrier? Animal traders in Guangzhou are handing over masked palm civets for slaughter.

India targets local HIV strain in test of AIDS vaccine

K. S. Jayaraman, New Delhi

India is gearing up to test an AIDS vaccine targeted at its most prevalent strain of HIV for the first time. Safety trials of the vaccine are expected to begin this summer.

With 60,000 AIDS deaths confirmed so far and an estimated 4.5 million infections, India is second only to South Africa as the country hardest hit by AIDS. About 30 AIDS vaccine trials are already under way worldwide, and the first Indian trial may go some way to address concerns that India has been slow to respond to the epidemic.

"There have been no vaccine trials in India, and that's obviously a real worry," says Seth Berkley, president of the International AIDS Vaccine Initiative (IAVI), a non-profit organization based in New York that is helping to organize the trial. "This is critically important, not only to India, but to the global effort to find a vaccine."

The National AIDS Research Institute in Pune in western India is hoping to conduct the phase I trial, which is only intended to establish the safety of the vaccine, on 40 healthy volunteers in June.

India has in the past rejected plans to test imported vaccines as they did not target HIV subtype C, which is prevalent in India. Early vaccine candidates had been aimed instead at subtype B, which is prevalent in America and Europe.

Since 2000, the Indian government has been working with IAVI to develop and evaluate an India-specific AIDS vaccine. "This vaccine is now ready," says Sekhar Chakrabarti, a scientist at the Indian Council for Medical Research who is involved in its development.

The vaccine to be tested is a live attenuated modified vaccinia Ankara virus containing the coding sequence of six HIV-1 subtype C genes isolated from an Indian strain. For the phase I trial, it will be produced by Therion Biologics, a biotech company based in Cambridge, Massachusetts.

Some AIDS groups have attacked the plan for the safety trial, which has yet to be approved by regulators. "We should abandon the plan, as there is no track record of success by previous candidate vaccines," says Ishwar Gilada, secretary general of the Peoples Health Organization, an AIDS-prevention group in Mumbai. But Chakrabarti says that the vaccine candidate has a greater chance of success than older vaccines.

Antibodies to SARS-like virus hint at repeated infections

Helen Pearson

A virus similar to that responsible for severe acute respiratory syndrome (SARS) infected people in Hong Kong 18 months before SARS reared its head, a recent study says. But some experts say the result is tentative and needs to be confirmed by larger studies.

In a finding published online on 2 January, a team led by Bo Jian Zheng, a microbiologist at the University of Hong Kong, analysed nearly 940 blood samples that had originally been collected in 2001 for a research project on hepatitis B (B. J. Zheng *et al.* www.cdc.gov/ncidod/eid/vol10no2/03-0533.htm; 2004). SARS first arose in China in November 2002, and public-health experts identified it and alerted the world in March 2003.

In the study, 17 samples carried antibodies either to the human SARS coronavirus, to a closely related animal coronavirus, or to both. Most of those who tested positive had a stronger response to the animal version.

The presence of such antibodies indicates that the Hong Kong inhabitants were exposed to SARS or to one of its animal cousins long before the disease was known or named. Such infections "probably came from a market animal or an environment contaminated by animal viruses," says Zheng.

Experts believe that the 2002 epidemic started when a coronavirus jumped from animals to humans, probably in the live-animal markets of Guangdong province, China. Researchers have found that several market animals and pets, including palm civets and ferrets, can harbour closely related viruses.

The result suggests that the leap from

animals to humans was not a one-off event, infectious-disease experts say. A whole family of SARS-like viruses may have been lurking in animals for years, occasionally spilling over into people, says virologist Albert Osterhaus of Erasmus University in Rotterdam, the Netherlands. "There must be other viruses out there," he adds.

In most cases, these coronaviruses probably infect human cells only weakly, perhaps triggering a passing sniffle. Zheng believes that the SARS virus mutated in animals or humans so it can readily jump between people, generating the potentially fatal infection.

The fact that at least one confirmed case of SARS in Guangdong has followed so hot on the heels of last year's outbreak supports the idea that SARS or related viruses can readily move between animals and humans, Zheng believes.

Zheng looked for SARS antibodies in a relatively large, healthy population, and says that similar investigations are under way in China. A 2003 study showed that 40% of people trading in wild animals in the Chinese markets also carried antibodies to SARS (Y. Guan *et al.* *Science* **302**, 276–278; 2003).

Some experts caution that the results are not conclusive, however. "I'd like to have seen ten times the number tested," says Christian Drosten of the Bernhard Nocht Institute for Tropical Medicine in Hamburg, Germany.

The researchers now hope to analyse blood from 10,000 healthy people for SARS antibodies. Such studies to flag up hotspots where many people carry antibodies might help scientists to track down the animal source of SARS, says Osterhaus.

Neglected diseases brought in from the cold

Alison Abbott, Munich

The European Commission is launching an ambitious programme to study a group of rare, mostly untreatable, genetic diseases that have so far attracted little attention from research agencies.

The programme, called EUmito-combat, will study diseases caused by defects in the thousand or so genes that regulate the function of mitochondria. These subcellular structures provide the body's cells with energy by breaking down sugars and fatty acids.

Malfunction of a mitochondrial gene occurs in around one in 10,000 births. About half of these lead to a syndrome known as Leigh's disease, in which muscle weakness and brain degeneration usually result in death within five years.

The aim of the programme, which will support a consortium of 21 research groups from nine European countries, is to develop treatment strategies for these inborn disorders. But it could also have spin-offs for more common diseases such as cancer and Parkinson's disease, in which mitochondria may also be implicated.

Jan Smeitink, a consortium leader who runs a centre for mitochondrial disorders at the University of Nijmegen in the



Power failure: defects in the genes that regulate mitochondria, the cell's energy generators, can result in fatal diseases.

Netherlands, says the backing is "the only way that progress can be made".

The number of patients in each category of mitochondrial disease is too small to excite much interest from pharmaceutical companies. And existing sources of public funds often require researchers to pursue short-term goals, which Smeitink says "is not the way to work".

Researchers in the new programme will be expected to work closely with patients. And despite the relative rareness of the conditions,

these are not in short supply: "Our centre investigates 300–500 patients a year from the Netherlands and Germany," says Smeitink. He adds that the programme will help to organize all the basic research being carried out on the disease in various laboratories around Europe.

This basic research involves studying mitochondrial genes, including those encoding proteins of the main mitochondrial energy circuit known as oxidative phosphorylation, or oxphos. This circuit produces adenosine triphosphate (ATP), the chemical energy currency of the cell. "We'll be looking systematically for defects and their cell-biological consequences," says Smeitink.

The programme also aims to develop diagnostic tools, and to use improved information on the basic biology of mitochondrial disorders to design treatment strategies—including gene therapy.

At present, only one type of mitochondrial disorder can be treated. It is caused by a very rare defect in the synthesis of an oxphos protein called coenzyme Q. Symptoms can be ameliorated, and deterioration delayed, by oral administration of this protein. But in most disorders, simple replacement of the damaged component does not work. ■

National park's sale of creationist book draws geologists' ire

Rex Dalton, San Diego

A slim volume of creationist views on how the Grand Canyon formed has sparked a legal review of books on sale in US national parks.

Lawyers for the National Park Service (NPS) began the review late last year after leading scientists objected to the sale of the creationist book, *Grand Canyon: A Different View*, in the bookstore at the Grand Canyon National Park in Arizona.

The review is intended to help the NPS to develop a policy for the sale of such material. "The book has raised issues much broader than this book and the canyon," explains David Barna, a spokesman for the NPS.

The beautifully photographed book was compiled by Tom Vail, a river guide with no scientific training. It features a collection of essays by 23 creationists who argue that the towering canyon walls are a record of the six days of creation which, they say, took place about 6,000 years ago. Most geologists say that the Colorado River cut out the canyon

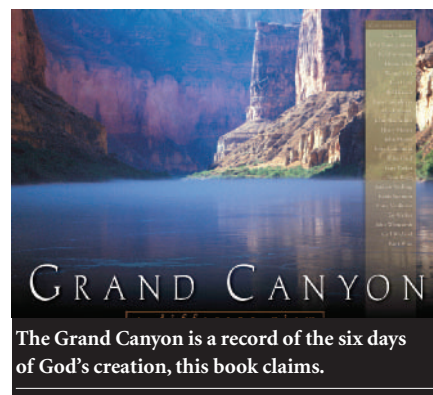
between 4 million and 6 million years ago, exposing 1.8 billion years' worth of geological formations.

Wilfred Elders, a geologist at the University of California, Riverside, who first complained about the book's sale at the park, called the volume a collection of "absurdities" that scientifically is "an extraordinary failure".

Complaints from Elders and others caused seven leading Earth-science organizations to write to the NPS on 16 December asking for the book to be removed from the shop—or at least for it to be separated from legitimate scientific texts with which it had been placed last August.

"The book is not about geology, but rather advances a narrow religious view," says the letter signed by presidents of the organizations, which included the Geological Society of America and the American Geophysical Union.

The 104-page book has subsequently



been moved from the science section, but remains on sale in the bookstore. Nearly 300 copies have been sold, according to the bookstore.

Vail says that an alternative to evolutionary science should be offered to members of the public visiting the canyon. "Who is to say whose material should be or shouldn't be in the bookstore?" he asks. That's the tricky question that the NPS review will seek to answer, as it weighs issues such as the display of sound science, the right to free speech and the avoidance of censorship charges. ■



Branching out: a 'sandpit' exercise on blood-vessel formation led to some innovative research projects.

Ethics accusations spark rapid reaction from NIH chief

Erika Check, Washington

The US National Institutes of Health (NIH) will act immediately to address concerns that its ethics standards are not rigorous enough, its director, Elias Zerhouni, has told Congress. But life-sciences lobbyists are already worried that the ethics issue is hurting the agency's budget outlook for 2005.

Concerns about ethics at the NIH were raised last month in the *Los Angeles Times* (see *Nature* 426, 741; 2003). The newspaper alleged that some of the agency's employees have been inappropriately influenced by consulting arrangements that were deliberately shielded from public view.

Billy Tauzin (Republican, Louisiana), chair of the energy and commerce committee in the House of Representatives, which oversees the NIH, had given Zerhouni until 8 January to respond to the allegations.

In letter sent in response to Tauzin, Zerhouni outlined a four-point action plan to address the concerns. He said that the NIH will review all external payments received by its employees since 1999. He also pledged to set up an ethics advisory committee inside the agency, as well as an outside panel of experts to advise the NIH on its ethics policies and practices. Finally, he pledged to review the policies that dictate when and how NIH employees disclose their consulting relationships to the public, and to "act to increase appropriate financial disclosure".

"Our mission is too important to the public health of the Nation to have it undermined by any real or perceived conflicts of interest," Zerhouni wrote.

Additionally, on 5 January, the NIH issued rules pertaining to conflicts of interest among scientists who serve on its scientific review panels. The rules largely articulate existing practice, observers of the agency say, and are not expected to make much difference to researchers who review NIH grant applications.

Officials at scientific societies praised Zerhouni's rapid response to the criticism. But they say that the issue has come at an awkward time for the agency, which faces an uphill fight for funding this year. President George Bush is expected to propose an increase of about 2% in the NIH's \$27-billion budget when he sets out his 2005 budget early next month — much less than the agency has enjoyed in recent years.

Sandpit initiative digs deep to bring disciplines together

Jim Giles, Dundee

A British research agency is trying out a fresh approach to creating interdisciplinary projects. It throws researchers from different fields into a 'sandpit' for a week with the promise of £1 million (US\$1.8 million) in support for the best ideas to come out of it.

The Engineering and Physical Sciences Research Council (EPSRC) tried out the concept for the first time in Dundee, Scotland, last month — and participants pronounced it an instant success. The topic was angiogenesis — the formation of new blood vessels — and the outcome was two proposed collaborations that the agency says it is likely to support.

The exercise wasn't without its pitfalls, however, as researchers struggled to get inside one another's heads. "I just don't understand the language they use," sighed Ana Schor, a cell biologist at the University of Dundee, after hearing a talk by a mathematician.

Although multidisciplinary workshops are commonplace, EPSRC officials say that their allocation of £1 million to support proposals from the sandpit was a new departure, which they hope to build on. "It worked," says John King, a mathematician at the University of Nottingham who helped to organize the meeting. "I'd like to see the approach applied to funding other subjects."

About 40 mathematicians, engineers, biologists, clinicians and others attended the session, and differences between their use of terminology and methodology surfaced early on. Mathematicians, for example, complained that the biologists wanted to build

models containing too many variables, whereas the biologists were frustrated at what they saw as the mathematicians' inability to explain clearly just what their models could achieve.

Some discussions went round in circles for hours as participants failed to settle on an aspect of the problem that could hold everybody's attention. "The social dynamics are as interesting as the science," says Oliver Jensen, an applied mathematician at the University of Nottingham. When proposals for grants did start to emerge, tensions rose again as different ideas competed for the prize on offer.

Two of these eventually won out. Researchers from Imperial College London and the universities of Nottingham, Oxford and Dundee are expected to get £600,000 to develop a kind of three-dimensional Petri dish. The dish will be used to examine how molecules involved in blood-vessel formation cause cells to release other compounds, for example, and to migrate. In the long term, they hope to simulate the formation of complete blood vessels in the device.

Another project, led by Christopher Mitchell, an expert in angiogenesis at the University of Nottingham, is set to get £400,000 to use new imaging technology to observe the formation of blood vessels in live mice. Both projects are expected to be funded after formal peer review.

"The meeting certainly generated ideas," says Robert Keatch, a microengineer from the University of Dundee. "It was more intense than a conference because we had to come up with a research proposal in five days." ■

UK slammed for weak support of science in poor countries

London The Royal Society this week criticized Britain's support for science in developing nations as uncoordinated and lacking long-term vision.

In a report to the House of Commons' Science and Technology Select Committee, the society described the government department that funds such initiatives as short on in-house scientific expertise and isolated from other parts of government. As a result, the report says, the Department for International Development is poorly placed to identify and support the science in poorer countries most relevant to their needs.

The department tends to focus on projects with direct benefits, such as boosting vitamin A levels in sweet potatoes in Uganda. This means it misses other projects that could aid several countries at once — such as satellite-based monitoring of rainfall in Africa, which would assist famine warnings and crop-yield assessments, the report says. It calls for a chief scientist, backed by a science team, to be appointed at the department.

Plan to rebury skeletons meets with grave response

London The director of the Museum of London has sent shockwaves through Britain's archaeological community by suggesting that much of his museum's collection of 17,000 human skeletons be given a Christian burial.

Jack Lohman told *The Times* newspaper last week that about 70% of the human remains excavated from sites in London appeared to have originally had Christian burials. He proposed that these bones should be studied and then reburied.

Museum officials say that Lohman's opinions do not reflect museum policy, but admit that they are considering the future of the collection and that reburial is an option.

Lohman's comments frustrate others who are fighting to retain control of their collections of human remains. "There are great medical benefits to humanity from skeletal collections," says Neil Chalmers, director of the Natural History Museum in London. "These must be weighed against the ethical issues."

Academy set up to boost Texan science

Washington Texas has founded its own Academy of Science, Engineering and Medicine in a bid to promote science and technology there. It also hopes to attract more research money and prestige to the

Sweet taste of success for honeybee research

Washington Genomicists have something new to buzz about: the honeybee genome, the first draft of which was deposited in public databases last week. The honeybee (*Apis mellifera*) is a major pollinator of crops, and is also a social insect, making the genome important for both agriculture and basic research, according to the sequencing team from Baylor College of Medicine in Houston, Texas.



second-most-populous US state, which ranks only fifth in federal research funding.

The academy was launched at a two-day conference in San Antonio, organized by Senator Kay Bailey Hutchison (Republican, Texas). Membership consists of the state's 11 Nobel prizewinners and all 200 Texas-based members of the National Academy of Sciences, the National Academy of Engineering and the Institute of Medicine. Local industry chiefs from high-tech companies such as Texas Instruments will act as advisory members.

Nobel laureates Richard Smalley, at Rice University, Houston, and Michael Brown of the University of Texas Southwestern Medical Center at Dallas will co-chair the academy's scientific advisory committee.



Richard Smalley: co-chair of Texas's Academy of Science, Engineering and Medicine.

University bids for chance to host nuclear waste

Tokyo Academics at Seoul National University in South Korea are hoping to bring a much-shunned nuclear-waste site to their workplace. A proposal, signed by 63 of the university's professors, advocates setting up the waste site in Gwanak, a mountainous area at the southern end of the campus.

The location of the waste site — South Korea's first — has been under discussion for 18 years. But in December 2003, the government opened the project up to bids

from around the country. It plans to decide on a site this year to begin construction of the facility in 2007.

The academics, led by physicist Chang-Sun Kang, say that they want to demonstrate their confidence in the safety of nuclear energy, but the plan has already met with protests from students and residents of Gwanak district. University officials are set to decide whether to go ahead with the proposal this week.

Germany aims high to create elite university

Munich Chancellor Gerhard Schröder announced plans last week to create at least one elite university to compete with top institutions in the United States and Britain.

Speaking at the Social Democrat's New Year retreat in Weimar, Schröder highlighted research and education as key areas of government expenditure, and revealed plans to give funding priority to one or a few selected universities.

Germany's hundred or so research universities are underperforming compared with the world elite. According to a recent ranking by the University of Shanghai's Institute of Higher Education, the best German institution, the University of Munich, is ranked only 48th in the world.

Reaction to the plan has been mixed. Members of the Green coalition expressed concern that the rest of Germany's already underfinanced higher-education system would suffer. But the heads of Germany's top universities have welcomed the scheme.

Correction

A News item erroneously implied that Michele Cargill of Celera Diagnostics in Alameda, California, and her colleagues used the draft chimp genome sequence generated by a public consortium to compare more than 7,500 chimp, human and mouse genes (*Nature* 426, 746; 2003). In fact, the team used its own chimp sequence data.

As one door closes...

Immigration controls introduced under the 'war on terror' are restricting the flow of foreign researchers into the United States. With other countries moving in on this pool of talent, will the balance of scientific power shift?

Zhang is a fifth-year chemistry graduate student at the University of Wisconsin at Madison. He is hard-working, popular with his colleagues, and should be on the threshold of a rewarding future in science. Yet a 2002 visit to Zhang's native China nearly derailed that career. He is so scarred by the experience that he agreed to be interviewed only on condition that his real name was not used in this article.

Nature's reporters are used to Chinese scientists requesting anonymity before speaking openly on controversial issues. But Zhang is not worried about the attitude of the government in Beijing. Rather, he is wary of consular officers, FBI operatives and other officials of the US federal government who seem to regard him as a potential terrorist, rather than a valuable member of their country's scientific workforce.

Zhang's nightmare began in January 2002, when he left Madison to spend the Chinese New Year with his friends and family. Zhang knew that immigration controls had been tightened up since the terrorist attacks of the previous September, and sought advice from his university about how to avoid any problems getting back into the United States. He carried with him proof of enrolment, details of the courses he had

taken, a letter from his department and government forms confirming his immigration status — which he assumed would allow him to get his student visa renewed. "I did all that I could have done," Zhang says.

But when he went to the nearest US consulate for an interview, Zhang was told he would have to wait. His particular field of study overlapped with a 'watch list' of technologies of potential interest to terrorists that had been supplied to consular officials. This meant that his application would have to undergo an interagency security review, involving security officials from agencies including the FBI and the Department of State.

Days stretched into weeks, then months, with no news of progress with his application. Eventually, Zhang found himself working in the office of a shipping company to make ends meet, while his colleagues continued their research without him. Because he didn't know when he was going to return, he was forced to continue paying rent on his apartment in Madison. Zhang finally received his visa in September 2002, leaving him hopelessly behind with his PhD studies. "My whole plan for graduation has been postponed," he says.

Zhang is not an anomaly. "There have

been enormous problems," says John Wright, who chairs the University of Wisconsin's chemistry department. Most of the students and postdocs whose applications to enter the United States have been questioned have eventually been let in. But Wright frets that the new immigration rules will deter future applications, weakening his department, which is currently considered among the best in the world. "The quality of research will decrease," he says.

"People who were thinking about coming to the United States for graduate school are now thinking twice."

Many US researchers and university officials share Wright's concerns. The United States is a nation of immigrants, and nowhere is this more evident than in the

country's research labs. Strip away the legions of foreign PhD students, postdocs and tenure-track researchers, and the behemoth that is the US scientific enterprise would look much less impressive (see figure, overleaf). What's more, in recent years, other countries have realized the value of attracting the best of the world's young researchers, and have started taking steps to compete more effectively in this marketplace (see 'You're welcome', below).

So will the United States' draconian response to the terrorist threat cause a funda-

You're welcome

At this time of year, Britain and Australia could be on different planets. In London, Cambridge and Edinburgh, it's cold, dark and almost unrelentingly wet; in Sydney, Melbourne and Perth, it's time to slap on the sunscreen and head for the beach.

But for many young scientists in Asia and elsewhere, these two countries have something important in common — they're among the most attractive destinations, now that delays in obtaining visas are restricting the flow of students and researchers into the United States.

Britain and Australia may be in the vanguard of the 'war on terror', but they don't share the US administration's view that foreign scientists should be viewed as potential terrorists. To government officials and university administrators alike, they're seen as valuable assets. This is especially true of fee-paying students, who bring a welcome influx of funding into the two countries' universities.

Australian universities have been particularly entrepreneurial in their efforts to attract students from Asia and the Indian subcontinent. "About five years ago, Australia wouldn't have entered



A place in the sun: Australia has made itself into an attractive destination for foreign students.

the consciousness of Indian students looking to study abroad," says Rohan Arthur, an Indian PhD student who came to James Cook University in Townsville, northern Queensland, to study coral ecology. "But now it's a big option."

That's largely because of the efforts made by an organization called IDP Education Australia. It was established in 1969, primarily as an outreach

programme to provide agricultural training for students from the developing world. Today, it has offices all over the world, which make strenuous efforts to attract students to Australian universities — IDP staff also assist visiting students in obtaining visas and accommodation.

As a result of these efforts, and similar moves by individual universities, the number of foreign



Fighting back: security clamp-downs on foreign researchers and students have sparked widespread protests at US universities.

students entering Australia was rising rapidly even before the current US immigration restrictions began to bite — averaging a brisk 13% per year over the 1990s, according to government figures. “I don’t think Australia depends on negative sentiment towards the United States for its growth,” says Anthony Pollock, vice-president international at Monash University in Melbourne.

But obstructive US visa policies have clearly accelerated the trend. Since 2001, the number of foreign science students enrolled at Australian universities has shot up by 32%. And according to IDP figures, the number of students in all disciplines coming from China jumped by 23% last year; for Indian students, the increase was a phenomenal 34%.

For young Asian scientists, Britain is less obviously appealing. “It’s very rainy,” observes Yingjie Liu, a Chinese pharmacologist who is studying towards a PhD at the University of Cambridge. The food isn’t much to her liking either. But Cambridge and other leading British universities have an excellent international

reputation, and are now exploiting this to attract Chinese and other foreign students in record numbers.

Every year, the Cambridge Chinese Students and Scholars Association holds a party to welcome new members. If present trends continue, they will have to move to a bigger venue. “Last year we had around 50 people at the party,” says Changxin Wu, the society’s president, who is studying for a PhD in veterinary science. “This year we had 120.”

The number of foreign students at Cambridge has been growing steadily over the past decade. Cambridge and other British universities have an active presence at international education fairs, and British embassies are also helping with the recruitment drive.

Universities in Britain and Australia have another important advantage — their native tongue is English, the lingua franca of science. For the same reason, Canada and Singapore are also benefiting from current US visa policies. But countries that don’t have this inbuilt advantage are competing with increasing effectiveness in

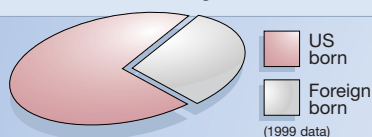
the international market for students and young researchers.

Officials in Germany, for instance, have long sought to dispel the popular notion that their country is hostile to foreigners. The government provides fellowships for visiting scientists and generous support for its universities’ presence at international career fairs. The charitable Alexander von Humboldt Foundation, which similarly provides fellowships for foreign scientists, even gives an annual award to the ‘friendliest’ of Germany’s immigration offices, which are responsible for issuing documents such as work permits.

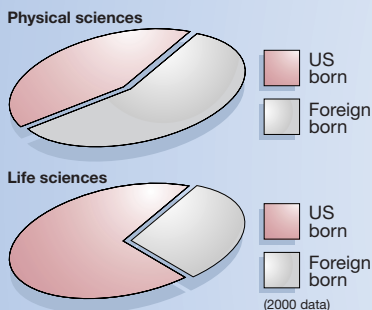
These policies are bearing fruit. Since 1999, the number of foreign students at German universities has increased from 113,000 to almost 200,000. Visiting researchers are also flocking to the 80 institutes of the Max Planck Society: since 1997, the number of scientists from the former Eastern bloc has more than doubled; over the same period, the number of guest scientists from China and India has tripled and almost quadrupled, respectively.

A foreign affair — immigrants and US science

PhD scientists working in the US

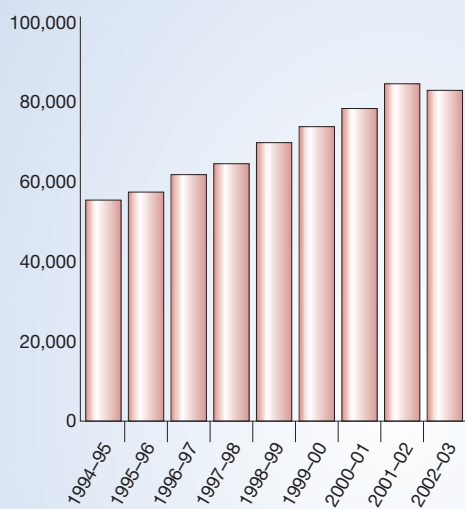


New PhDs awarded in the US



Source: National Science Foundation (NSF)

Total number of visiting scholars to the US by year



Foreign-born US residents with science or engineering doctorates

Place of birth	Number
China	37,900
India	30,100
United Kingdom	13,100
Taiwan	10,900
Canada	8,400
Germany	7,200
Iran	4,800
Former Soviet Union	4,600
South Korea	4,500
Philippines	3,400
Poland	3,200
Japan	2,800
Argentina	2,700
Other foreign-born	58,400

Source: NSF

mental shift in the international movement of researchers — and perhaps even alter the global balance of scientific power? It's difficult to say, because attaching firm numbers to such trends is all but impossible. Scientists travel to the United States on a wide variety of visa types, depending on the purpose and length of their stay. And because they make up a tiny proportion of the total number of foreigners entering the coun-

Wrapped in red tape: some foreigners struggle to enter the United States and face security checks even after arrival (below).

try each year, even a major decline would fail to show up in overall visa statistics. Data collected in different countries are also hard to compare: many nations don't separate visiting scientists from researchers in the humanities and other disciplines; some consider students separately from postdoctoral researchers, whereas others lump them all together.

"One of the great problems in dealing with this issue is that you get tons of anecdotes, but it is difficult to get data," says

Norman Neureiter, who served as science adviser to US Secretary of State Colin Powell for three years until September 2003. *Nature's* enquiries reinforce Neureiter's view of the anecdotal evidence. Our reporters found dozens of examples of scientists at every level who have experienced problems entering the United States. And in some cases, they found researchers now looking for work in countries such as Australia, Britain and Canada, rather than enduring the US immigration process.

Out in the cold

The sketchy data available suggest that such anecdotes illustrate a widespread problem — and that this is particularly acute for postdoctoral researchers in the 'hard' sciences and engineering. In November, for instance, the Association of International Educators, an organization based in Washington DC that promotes scholarly exchange worldwide, released a survey of more than 300 US colleges and universities. The survey revealed that the number of students whose start dates were delayed by visa problems was 48% higher in 2003 than at the start of the previous academic year; for 'scholars' — a broad category dominated by young postdoctoral researchers — the increase was 76%. More than three-quarters of the delayed students were in the physical sciences, biological sciences or engineering; among the scholars, these disciplines accounted for 93% of those who experienced significant delays.

Other surveys paint a similarly bleak picture. Last July, the American Institute of Physics reported that nearly a quarter of foreign students who applied to study towards a PhD in physics in the United States in 2002 were initially denied a visa. The number of



foreign researchers working at the five largest institutes on the National Institutes of Health campus in Bethesda, Maryland, declined in 2003 for the first time in the nine years over which records have been kept. Most strikingly, the total number of visiting scholars in the United States declined in the 2002–03 academic year for the first time in almost a decade (see figure, opposite).

For some observers, these statistics are enough to set off alarm bells about the future health of US science. “We’re at a critical juncture now, and I think everybody senses it,” says Irving Lerch, director of international affairs with the American Physical Society in College Park, Maryland. Although the likely consequences of the visa delays remain a matter of debate, their main cause is clear — new security procedures introduced following the terrorist attacks of 11 September 2001.

In the immediate aftermath of those events, the state department began expanding its ‘Technology Alert List’, designed to prevent dangerous technologies getting into the hands of terrorists or hostile states. It is now classified, but a version issued in August 2002 contained roughly 150 items, including such broad labels as ‘microbiology’, and common pieces of lab equipment such as low-energy lasers. So if you work on, say, infectious disease, or use relatively innocuous devices that have found their way onto the state department’s list, your application to enter the United States is likely to be referred to the FBI and other federal agencies for a security review.

Singled out

Scientists from China have borne the brunt of the new policy — even though its nationals have never been implicated in terrorism against US targets. The survey by the Association of International Educators, for instance, found that more than a third of all visiting students whose entry to the United States was delayed were from China. In part, the large number of Chinese who have been affected by the new restrictions reflects the fact that they make up the biggest single group of foreign scientists seeking employment or education in the United States. But some Chinese researchers, who point out that the current US administration was pursuing an aggressive policy towards their country even before the 2001 terror attacks, believe that they are being singled out for harsh treatment (see ‘We are not the enemy’, right).

Meanwhile, for researchers from countries such as Iran, and several others in the Middle East, security reviews have become an almost insurmountable barrier (see

We are not the enemy

You don’t need to visit Beijing or Shanghai to witness the pain of young Chinese scientists who have found themselves locked out of the United States, frustrated in their attempts to obtain the necessary visa. It’s laid bare on the Internet for all to see, in discussion forums run by associations of Chinese scholars.

The New York-based Columbia University Chinese Students and Scholars Association, for instance, runs an online bulletin board on which delayed visa applicants, who call themselves ‘checkees’, record their ups and downs, and trade advice on how to get their lives back in gear. “What happened is totally a nightmare. Waiting helplessly at home and watching my study being ruined is so disappointing and frustrating,” states the entry from Guoyuan Liu, a chemistry graduate student who went home to get married in December 2002 and waited 12 months for a visa to return to his studies. The delay cost him his financial support at Vanderbilt University in Nashville, Tennessee, forcing him to transfer to the PhD programme at the University of Maryland. On a similar site run from Tsinghua University in Beijing, one checkee who finally gained a visa in late October 2003 issues a stark warning to those who might consider a similar trip home: “Exercise extreme caution.”

On another page of its website, the Columbia association collates information on checkees trying to enter institutions across the United States. Last updated on 9 September, the list covers 679 visa applications, 280 of which had been approved and 399 were still pending. There had been no outright rejections, but the average delay before approval was 154 days.

Some Chinese researchers suspect that concerns about security are being used as an excuse to discriminate against them. Many remember the aggressive stance adopted towards China by President George W. Bush in the early days of his administration, before the terrorist attacks of 11 September 2001. Wendy White, who directs the Board on International Scientific Organizations at the National Academies in Washington DC, suspects that concerns about industrial espionage may explain some of the heightened scrutiny facing Chinese researchers. She adds that sources in the FBI have told her that the delays being faced by Chinese visa applicants often reflect genuine difficulties in conducting security checks caused by the similarity between many Chinese names.

Qikun Xue, a materials scientist at the Chinese Academy of Sciences’ Institute of Physics in Beijing, was once a regular visitor to the United States. But since 2002, visa difficulties have caused him to miss the American Physical Society’s March 2002 meeting and prevented a collaborative project with a US colleague. “The experience gives me the feeling that the delay in issuing the visa might be done intentionally,” he says.

Whatever the reason, the delays do seem to be having an impact on the flow of Chinese scientists into the United States. In the 2002 academic year, the number of visiting Chinese scholars declined for the first time in a decade; the number of US visas issued to Chinese students is also down 24% since 2001. It’s not difficult to see where these young scientists are going — in Australia, Britain and Canada, the number of Chinese students has roughly doubled since 2000.

For the United States, the consequences of this trend could be severe. About one in five of the foreign scientists currently working in the country are from China — and if you doubt their importance, take a quick scan for Chinese names on papers from US labs in the top journals in your field. “There’s going to be a long-term impact on the United States’ ability to attract the best students and researchers,” says Bing Su, an evolutionary geneticist at the University of Cincinnati, Ohio, who also has a position at the Chinese Academy of Sciences’ Institute of Zoology in Kunming.

♦ bbs.columbiachinese.org/viewforum.php?f=8
♦ www.columbiachinese.org/checkee_list.html
♦ www.smth.org/bbsdpc.php?board=Visa
♦ www.smth.org/bbsdpc.php?board=VisaCheck



No entry: a Chinese student displays his visa rejection stamps from the United States.



On the record: new rules dictate that all foreigners who need a US visa must be photographed and have their fingerprints taken when they cross the border.

'Never apply for a US visa again!' opposite). Because the US government sees Iran as a sponsor of terrorism, its scientists cannot enter the United States without undergoing an interagency review. Even senior Iranian officials with longstanding ties to the US scientific community have been unable to attend major conferences. "I have had many invitations, but I had to say no," says Reza Mansouri, a physicist and the deputy of research at the Iranian Ministry of Science, Research and Technology in Tehran.

Scientists from other countries need not face a full security review, even if their work appears on the state department's watch list. But a memo sent in August 2002 along with a revised version of the list ensnared many scientists who expected to sail through the immigration process. The paper instructed that, "when in doubt", consular officers should send applications to the state department's headquarters in Washington DC. As the consular staff involved were mostly inexperienced, they were in doubt all too often. The resulting backlogs caused delays of up to a year.

The memo is still causing problems. The state department claims that more than 80%

of cases referred to Washington are dealt with in 30 days. But Wendy White, who directs the Board on International Scientific Organizations at the US National Academies, disputes this figure. "For the scientists we hear from, the average wait time is still over five months," she says.

Delays were exacerbated last July by a new rule requiring virtually all visa applicants to be interviewed face-to-face by a consular officer. Most scientists were already being pulled into US embassies for interviews, but they suddenly found themselves part of a much longer queue. When Thomas Brunold, an assistant professor in chemistry at the University of Wisconsin, went home to Switzerland for a short visit last June, he had to wait for three months to get an interview to renew his US visa. "I told them I had a research group of nine people to run," Brunold says. But his pleas fell on deaf ears, and the resulting delay cost Brunold a month's salary.

For many researchers, the most frustrating thing about the new immigration requirements is their inconsistency. As a result, some visa applications shoot through the system whereas others are held up

for months. And when this happens, there is usually no explanation. "The transparency in the process is completely missing," says Olexei Motrunich, a Ukrainian physicist who has worked in the United States since 1994, but has been stranded in his home country since July, unable to take up a postdoctoral position at the University of California, Santa Barbara.

"I have been telling my relatives and friends how great America is; how one does not feel foreign in this country," says Motrunich. "Now I have to explain to the same people why, after more than eight years of doing science in the United States, I have a hard time receiving a visa to continue my work."

Number one no more

For some visiting scientists, the problems don't end at the US border. Catheryn Cotten, who directs the International Office at Duke University in Durham, North Carolina, says that foreign nationals are finding it more difficult than ever to secure social-security numbers, driver's licences and other essential documents. Mansouri adds that press reports of assaults against Iranian students at US universities are causing many of his country's young scientists to think instead about studying in Britain or Australia.

"You can't go to a large international scientific meeting without visas being the issue on everyone's mind. I think there's going to be a solidarity movement against the United States."
— Wendy White

Such comments are worrying organizations that strive to promote international scholarly exchange. "There's a perception that visas are too difficult to get and the United States is an unwelcoming place," says Victor Johnson, associate executive director for public policy at the Association of International Educators.

Not surprisingly, researchers and university administrators in other countries who are recruiting from the pool of scientists now experiencing problems entering the United States are quietly satisfied with the turn of events. Countries such as Australia, Britain and Canada were already increasing their intake of foreign students before US visa restrictions were imposed — and this trend has accelerated since then.

Perhaps even more significant is the

calibre of the students and researchers now considering destinations other than the United States. "I've had professors tell me that the quality of the Iranian students is phenomenal," says Amy Aldous, graduate-studies recruitment manager at the University of Waterloo in Ontario, Canada. Baowen Li, a theoretical physicist at the National University of Singapore, says that he is now seeing many more applications from China's elite universities. "The change is not in quantity but quality," says Li. "We have benefited a lot from the US policy."

But is this the start of a trend that could ultimately undermine the United States' leadership in science? Andreas Schleicher,

"I can understand that Americans are worried about their security, but I don't understand why people like me are a threat to their security." —

Zahra Fakhraai

who heads the Indicators and Analysis Division at the Organisation for Economic Co-operation and Development in Paris, argues that US dominance is so overwhelming that this is unlikely. "More than a quarter of all students studying abroad still travel to the United States,"

he says. But Neureiter, who has wrestled with the issues from inside the US administration, is not so sanguine. "I tend towards an apocalyptic view," he says.

How things unfold will depend on whether the visa delays experienced by visiting scientists represent teething troubles or a more lasting obstacle. State-department officials argue that they are now taking steps to improve the situation. New rules should let students jump to the front of the interview line so that they do not miss their start dates. And by March, a new computer system should connect embassies overseas directly to security agencies in the United States. The idea is to speed the interagency security reviews, preventing cases such as Motrunich's from getting stuck in limbo.

Still, the ongoing focus on security means it will be impossible to handle applications as quickly as they were dealt with before 2001. "I think the best we can do is to try to keep with our goal of processing all of the cases within a 30-day period," says Janice Jacobs, deputy assistant secretary of consular affairs at the state department.

Some US universities report that things do now seem to be getting back on track: at Duke, for instance, the number of foreign students studying the sciences rose once more in 2003, after two years of zero growth. But Neureiter is worried about the potential impact of a rule implemented last week that requires the fingerprinting of all visa applicants, and of another that will soon demand that students and visiting scholars pay a non-refundable fee of \$100. "You can't go to a large international scientific meeting without visas being the issue on everyone's mind," agrees White. "I think there's going to be a solidarity movement against the United States."

Back in Madison, Zhang is now applying for postdoctoral positions, while writing up his PhD thesis. Despite his experiences, he says that he would rather stay in the United States, where he knows the research community. "But people who were thinking about coming here for graduate school are thinking twice," he warns. While Zhang was in China working at the shipping company, he befriended his boss's family. The executive's two daughters were thinking of studying medicine. Last year, they began their courses in Britain.

Written and reported by Geoff Brumfiel, with David Cyranoski, Carina Dennis, Jim Giles, Hannah Hoag and Quirin Schiermeier.

Never apply for a US visa again!

Zahra Fakhraai (pictured) and Sina Valadkhan had it all worked out. In May 2002, they were planning to relocate from Tehran to Massachusetts, having each received an offer to study towards a PhD at a prestigious US institution. Fakhraai would conduct research in polymer physics at Boston University, while her husband, Valadkhan — a gold-medal winner at the 1996 International Physics Olympiad in Oslo, Norway — would study mechanical engineering across the Charles River in Cambridge, at the Massachusetts Institute of Technology.

Because the United States has no diplomatic representation in Iran, the pair travelled to the US embassy in Ankara, Turkey, to apply for the necessary visas. They took medical records, academic records from Tehran's Sharif University of Technology, photographs, passports and letters that declared their intent to return to Iran after their studies. But their applications were rejected on the spot.

Two weeks later the couple tried again and were rejected a second time. On this visit, Fakhraai claims they were laughed at and granted only brief interviews by US consular officers. Fakhraai says that officials dismissed her husband's documents as possible forgeries. "They told me that I didn't have enough evidence that I would go back to Iran,"

Fakhraai adds. When she pointed out a standing offer to resume her research at the Institute for Studies in Theoretical Physics and Mathematics in Tehran, Fakhraai says that the officer told her he didn't need to see it.

The couple persevered and tried again in June — this time at the US embassy in Dubai, in the United Arab Emirates. The end result was the same, says Fakhraai: their applications were rejected. "I can understand that Americans are worried about their security, but I don't understand why people like me are a threat to their security," says Fakhraai.

Discouraged, the pair returned to Tehran, deferred their admissions to the Massachusetts schools, and applied to the University of Waterloo in Ontario, Canada. The new offers arrived in October, and the couple returned to Dubai again. At the American embassy the officer rejected their visa applications once more. When Fakhraai asked why, she claims she was yelled at: "Never apply for a US visa again!"

She didn't. The couple returned home, and in November 2002, Fakhraai and Valadkhan were granted study visas from the Canadian government on the day of their application. One year into her studies at Waterloo, Fakhraai is pleased with her move to Canada. "You never feel like a foreigner," she says. "Living and studying here is one of the best experiences I've had."



Scandals stem from the low priority of peer review

Good refereeing should be recognized and rewarded, with help from the journals.

Sir—As part of the continuing debate over the future of scientific publications, I wish to make two comments and to advance a suggestion.

First, what is the difference between a published paper and a paper that has been uploaded by its authors onto some electronic archive? The difference is that one has been refereed, the other has not. So, refereeing is clearly at the core of the scientific publishing business.

Second, what are the calls on a scientist's time? These, in order, are: (1) to bring more money into the laboratory; (2) to write and publish more papers and

conference communications; (3) to write reports for the funding agencies; (4) to get some science done; (5) to teach and/or carry out administration; (6) to sit on committees that distribute the money; and (7) to referee papers.

It is a real temptation today for a referee to clear a paper off his or her desk by writing a casual and ill-considered report. If refereeing is to be done better than of late (and recent scandals involving famous laboratories are revealing), it must move up the list of priorities.

The best solution would be to include refereeing as a factor in assessment for

recruitment and promotion, and perhaps even for the allocation of funds. This requires a measure to be devised, which should probably be based on information provided by the quality scientific journals.

If scientific publishers wish to avoid embarrassing fiascos and to ensure that the publication process remains reliable, they would do well to address this issue. Otherwise, the core of their business is at risk.

Jean-Patrick Connerade

*Euroscience, 8 rue des Ecrivains,
F-76000 Strasbourg, France*

Eastern Europe: progress stifled by the old guard

Sir—Your Editorial "Eastern promise" (*Nature* 426, 369; 2003) argued that integrating several eastern European countries into the European Union (EU) might boost European science owing to their untapped human potential. You suggest that integration will be a short-term challenge for the EU, but will ultimately strengthen it.

As a scientist from Poland — one of the EU members-to-be — who has worked both in Western Europe and now in the United States, I must strongly disagree with your conclusion.

It is true that there is great human potential in eastern Europe. However, you seriously underestimate the power of the scientific establishment in those countries — ranks of professors promoted during communist times — to hinder progress. Educated, hard-working scientists flee whenever possible, either by abandoning science altogether or by emigrating, usually to the United States, because the job market is tough in the EU. Few are brave enough to fight to work in their homeland.

The only way to unleash any hidden human potential is by drastic reform of science and higher education in countries such as Poland. Sadly, there are no signs of change on the horizon. Western scientists rarely understand how science works in the east. In Poland it is hierarchical, immobile, hermetic and gerontocratic. Recognition comes from having a professorship, and the postdoctorate qualification called *habilitation*, not from publications in internationally recognized journals with high impact factors. A scientific career after PhD and *habilitation* depends on personal and political

connections. Future professors require a certain number of publications, so they publish worthless papers in countless Polish 'scientific' journals.

Once they have been promoted, the professors are no longer required to do any real research. Their titles are bestowed for life, and a head of department keeps that position until retirement. Professors usually work in the university where they completed their undergraduate, graduate and PhD studies, where everybody knows everybody else. Outsiders are rare and nepotism is common. Entire generations gain professorships because they are relatives or favourites of previous professors. Most research money is distributed by arbitrary administrative decisions, not as peer-reviewed grants.

Polish universities are ruled by democratic elections, but the scientific establishment is not interested in change. Some professors are creating the illusion of reform under the auspices of the president of Poland — but it is difficult to expect them to undermine their own existence.

Integrating eastern European science into the EU will do more harm than good unless the EU enforces real reforms in those countries. To thrive, science in eastern Europe must become part of the international scientific community: the *habilitations* and titular professorships must be abolished, scientific merit must be the only measure of an individual's qualifications, and money for research must be distributed by a competition among peer-reviewed applications.

I fear that Europe lacks the political will to modernize science. Even within the EU, anachronisms in several countries make their science less competitive than that of the United States.

One day I would like to work again in Europe, especially in Poland. But as it is

now, the heart of science beats on the other side of the Atlantic.

Cezary Wójcik

*University of Texas Southwestern Medical Center,
Department of Physiology, 5323 Harry Hines
Boulevard, Dallas, Texas 75390-9040, USA*

Eastern Europe needs a competitive atmosphere

Sir — Eastern European science is underfunded and lags behind the major players. However, as your Editorial suggests (*Nature* 426, 369; 2003), the great human potential in terms of qualified scientists and excellent students provides a solid basis for growth.

During my scientific career I have had the chance to work both in the European Union (EU) and in Slovakia, which is due to join the EU this year. In my view, the problem is not only financial, as there are well-funded laboratories in eastern Europe with state-of-the-art equipment. What is missing is the competitive atmosphere found in top-class research centres. Foreign scientists are very rare in eastern European countries, which make no significant effort to attract experienced scientists from abroad — unlike emerging economies in Asia (see, for example, *Nature* 420, 257; 2002).

One solution would be to establish new international research centres in eastern Europe. Apart from the International Institute of Molecular and Cell Biology in Warsaw (*Nature* 421, 471–472; 2003), genuine international research centres have been absent in these countries. Establishing such centres will help prevent the brain drain and will be an effective way to harness talent in eastern Europe.

Juraj Gregan

*IMP (Research Institute of Molecular Pathology),
Dr. Bohr-Gasse 7, A-1030 Vienna, Austria*

Warning of warming

How data and modelling led to predictions of climate change.

The Discovery of Global Warming

by Spencer R. Weart
Harvard University Press; 2003. 240 pp.
 \$24.95, £16.95, €23.10

Stephen H. Schneider

In the mid-1990s, as editor of the journal *Climatic Change*, I was sent a paper outlining the history of the international climate-change assessment process over the previous few decades. The Intergovernmental Panel on Climate Change (IPCC) had recently taken centre stage as the world's 'scientific consensus' on global warming. The article was written by a policy-studies graduate student who had rarely gone to any of the meetings he was reporting on, nor did he know most of the principals whose efforts he chronicled. I wondered how he could possibly get the events and flavour of this evolution of climate efforts remotely right.

Nevertheless, I sent the paper to two senior climate scientists, both of whom had been present at the creation of this evolution of international cooperation in climate assessment. If the student, Shardul Agrawala, could get past these two tough reviewers, he would deserve to be published. I was amazed to receive two rave reviews. "How can someone who never was there do such a good and fair job that it reminded me of parts of this history I had forgotten?" one wrote. The paper was published with a few minor corrections and is now a standard for the history of this effort.

When I was asked by *Nature* to review Spencer Weart's history of the discovery of global warming, from the nineteenth to the twenty-first century, I was once again sceptical. According to his biography, Weart, a historian of science working at the American Institute of Physics, was not remotely a player in the climate debate. But, remembering my experience with Shardul's article, I began to read his compact little account.

It didn't take long for me to share the amazement of the reviewers of Shardul's paper. Weart's account brought back many forgotten memories of the great scientists I admired in the 1970s and had largely forgotten, including Murray Mitchell, Gilbert Plass and Fritz Moller. He also recounted the discoveries and issues raised by many others with whom I have been, or still am, friends



Getting warm: Roger Revelle (above) and Murray Mitchell were among the first scientists to detect signs of climate change.

and colleagues. Roger Revelle was my interdisciplinary mentor for 20 years, so I read the account of the work of Roger and his colleague Hans Suess in the 1950s with particular interest. "If just one of these men had been possessed by just a little less curiosity, or a little less dedication to laborious thinking and calculation, decades more might have passed before the possibility of global warming was noticed."

It is amazing that in such a short book, Weart has managed to stuff in most of the truly important events and characters in the evolution of the global-warming issue. To be sure, I could think of many issues that he left out — such as the need for transient calculations and the pioneers of signal-to-noise ratios in the 1970s who helped to bring about better interpretations of the model results — but Weart manages to cover many of these in an excellent website produced as a companion to the book (www.aip.org/history/climate).

My research assistant, Janica Lane, who is helping me to edit my own website (<http://stephenschneider.stanford.edu>) on

the recent history and controversies surrounding climate change, has come across every possible website on climate imaginable, from exaggerated sites of 'deep ecology' non-governmental organizations to the shrill nonsense of 'enterprise institute' anti-warming sceptics and their political apologists in Congress and the business media. "It was such a pleasure to read such a fair and balanced account," she commented after checking out Weart's website.

In the final chapters of the book, Weart departs from his detached historian role and addresses the modern political problems of climate policy. I found the largely precautionary personal values expressed here to be fair. But best of all, he offers this justification — almost an apology — for his delving into recent events: "The closer an account of events moves toward the present, the less it can be called 'history', and the more it looks like something else — perhaps journalism. The special virtues we seek in a work of history, the long-view perspectives and objective analysis, dwindle. Writers and readers find it hard to pick out which recent developments will really matter in the long run. Worse, opinions about present-day controversies infect views of the recent past with special virulence. Therefore be wary: this concluding chapter can be only a preliminary sketch."

But I'm glad that Weart did take on the present political climate scene, because he concludes that although scepticism is essential for the health of science, the small cadre

of professional 'contrarian sceptics' have been ideological, shrill and way out of step with mainstream science. This historian got it right, both in the past and where the issue is going. I only wish that more of today's journalists and politicians were so careful and insightful.

It is the unwritten duty of a book reviewer to complain about something. So let me do it with full narcissism. In citing my first atmospheric-science paper in 1971, which suggested that aerosol cooling could dominate greenhouse-gas warming, Weart says that the "equations and data were rudimentary, and critics swiftly pointed out crippling flaws". He is right about the crippling flaws, but what I am most proud of was pointing most of them out first myself in a 1975 paper and in my book written with Lynne Mesirov, *The Genesis Strategy* (Plenum, 1976). Weart does note a 1992 chapter in which I predicted that an unambiguous greenhouse climate signal would emerge from the climatic noise around the end of the century, but I had first made this point in *The Genesis Strategy*. Given the IPCC's comments about the "discernible" impact of humans on climate in 1995, I am pretty proud of my 1976 crystal-ball gazing.

But these few personal complaints are trifles. This is a terrific book. For example, despite the polemics today from those who point to uncertainties in climate science as an excuse for inaction, when the usual proposed actions such as carbon taxes would hurt their ideological or clients' interests, Weart recognizes that science operates that way: "Scientists rarely label a proposed answer to a scientific question as 'true' or 'false', but rather consider how likely it is to be true. Normally a new body of data will shift opinion only in part, making the idea seem a bit more or less likely."

This is a clear statement of the bayesian or subjective probabilistic framework that is becoming the standard for complex assessments of problems such as climate change. I only wish more of my own colleagues were as epistemologically sophisticated about uncertainties and subjective probabilities as this historian (see *Nature* **418**, 476–478; 2002).

Perhaps the finest compliment I could give this book is to report that I intend to use it instead of my own book *Coevolution of Climate and Life* (Sierra Club Press, 1984) for my climate class. *The Discovery of Global Warming* is more up-to-date, better balanced historically, beautifully written and, not least important, short and to the point. I think the IPCC needs to enlist a few good historians like Weart for its next assessment. ■

Stephen H. Schneider is in the Department of Biological Sciences and co-directs the Center for Environmental Science and Policy, Stanford University, Stanford, California 94305-5020, USA.



Role model: physicist Ayse Erzan won the 2003 L'Oréal-UNESCO For Women in Science award.

The parenting gap

Women in Science: Career Processes and Outcomes

by Y. Xie & Kimberlee A. Shauman
Harvard University Press: 2003. 336 pp.
\$59.95, £38.95

Abigail J. Stewart and Danielle LaVaque-Manty

Do young women take fewer mathematics and science courses in high school than young men, leaving them less prepared and therefore less likely to major in science and engineering fields in college? Is a woman with a bachelor's degree in science and engineering more likely to have begun her college career as a science major, or on a non-science track? This book, ten years in the making, offers definitive and surprising answers to these and other long-standing questions about women in science.

Using an inventive approach to deal with the paucity of data, Yu Xie and Kimberlee Shauman examine the question of women's under-representation in science by combining a 'life-course perspective' with the statistical analysis of 17 nationally representative data sets.

The life-course perspective assumes that major transitions in people's lives are "age-dependent, interrelated, and contingent on (but not determined by) earlier experiences and societal forces". By contrast, the more familiar conceptualization of career trajectories in science and engineering is a "science pipeline". This pipeline is unidirectional: participants enter the pipeline by taking maths and science courses at school, and leak from it at various points when they stop pursuing coursework or careers in science.

And it is one-dimensional, regarding women's relationship to science in isolation from everything else.

The analysis of multiple data sets allows the authors to construct 'synthetic cohorts' of women, or 'hypothetical cohorts whose life history is constructed from real cohorts', whose career processes can be compared to those of synthetic cohorts of male counterparts. The composite portrait generated should reasonably represent the lifetime career trajectories of the population of women in science.

The care that the authors take with their empirical approach allows them to offer definitive answers to important questions. They find that although young men are twice as likely as women to enter college with the intention of majoring in science or engineering, this is not explained by gender differences in high-school maths achievement or coursework. The gender gap in mathematics achievement is small and has been declining, and girls not only take as many maths and science courses as boys, but also get significantly better grades in them.

More surprisingly, Xie and Shauman find that the majority of men who get baccalaureate degrees in science or engineering pursue those degrees throughout their college years, whereas most of the women who graduate in these fields enter science and engineering during college after starting on non-science tracks. This discovery complicates the unidirectional image of the leaky pipeline.

Several chapters in the book point to the role that having children plays in women's career trajectories. Married women with children are most likely to leave science and engineering after completing a degree. They are also less likely to be employed, promoted or geographically mobile than either their

male counterparts or women, whether married or single, who do not have children. Thus, Xie and Shauman believe that the gender gap in parenting responsibilities is a barrier to women's progress in science careers.

They offer few policy recommendations, but at least one seems new. Given the likelihood that a woman who leaves college with a science or engineering degree began her studies in a non-science field, educators need to figure out how to make studying science more attractive to women who are currently majoring in something else. Recruitment at the undergraduate level may be at least as important as retention.

Xie and Shauman's findings also provide further evidence for the idea that employers should embrace policies that increase both flexibility (such as job-sharing and flexi-time) and the availability of on-site childcare for working mothers.

It is important to note that the methodology that enables Xie and Shauman to provide us with definitive answers to some kinds of questions is a blunt instrument when it comes to others. For example, the authors are explicitly unable to address any possible school-level influences on young women's career plans, and cannot distinguish between physics, which currently attracts few women, and the biological sciences, in which women earn as many or more degrees than men. Nor can they offer insight into questions of institutional climate and practice and their effects, including effects on post-undergraduate leakage from science.

This is not to disparage the book for what it does not do — Xie and Shauman's careful research answers hard questions that have, in the past, seemed virtually unanswerable — but simply to note the limitations inherent in using the kind of data available to them. Their work should serve as a stimulus to further research applying equally careful and creative approaches to the many questions that remain. ■

Abigail J. Stewart is in the Department of Psychology and the Program in Women's Studies, and Danielle LaVaque-Manty is at the Institute for Research on Women and Gender, University of Michigan, Ann Arbor, Michigan 48109-1109, USA.

The descent of man

Adam's Curse: A Future Without Men

by Bryan Sykes

Bantam Press: 2003. 300 pp. £18.99

Jennifer A. Marshall Graves

I think 2003 must have been the Year of the Sex Chromosomes. On the heels of Steve Jones' Y and David Bainbridge's *The X in Sex* (both reviewed in *Nature* 423, 223; 2003) comes *Adam's Curse: A Future Without*

Wildlife in watercolours

The German artist and natural historian Maria Sibylla Merian (1647–1717) was a remarkable woman who, as a single mother, earned her living as an artist and travelled to South America in search of new specimens to paint. She came from a family of artists — her father was an engraver, and both her stepfather and her husband were painters.

Maria had a keen eye for nature and as a child kept silkworms so that she could record their development in her paintings of flowers and insects.

She published many books of her own, although the painting of coconut crabs shown here was one of the illustrations she made for Georg Eberhard Rumpf's book *D'Amboinsche Rariteitkammer*.

Many of Maria's original paintings were purchased by Tsar Peter the Great for his art museum in St Petersburg, Russia. They are now available to a wider audience in *Maria Sibylla Merian: The St Petersburg Watercolours* (Prestel, £55).

Mary Purton



Men (note the absence of a question mark).

After reading Bryan Sykes' delightful article on the history of the Sykes Y chromosome (*Am. J. Hum. Genet.* 66, 1417–1419; 2000) and his successful book *The Seven Daughters of Eve* (Norton, 2001), I looked forward to this book. I admire authors who can interest non-scientists in genetics — a vital skill if we are to cultivate an informed public to debate the manipulation of sex and reproduction.

Indeed, the book is fun to read — the writing style is lively, the images fresh and witty, the explanations of basic genetic principles apt and accurate, even inspired. Like *Y, Adam's Curse* centres on sexual conflict, here the war between the mother's mitochondrial genome and the father's Y chromosome. Sykes traces the spread of the Y chromosome in space and time, enriching the account with the history of Vikings, Polynesians and Genghis Khan.

The author's focus on his own family is a good device to explain how the Y chromosome gets around and to introduce the history of families and surnames, migrations and conquests. But the focus on Sykes and his family, Sykes' blood cells and the Sykes Y chromosome, then Sykes' ideas and finally Sykes' wild speculations, rather gives the impression that the entire field was explored single-handedly by Bryan Sykes, genetic supersleuth.

Of particular interest to me were the dire predictions of the imminent decay of the Y chromosome. Sykes calculates from the

frequency of Y mutations in men (can it really be high as 2%?) that the fertility of the whole human population will plummet within 125,000 years (upping the ante on my calculation of 9 million years). But does the disappearance of the Y chromosome, as Sykes avers, really mean the extinction of humankind unless we can dispense with the imprinting of at least 100 genes and embrace parthenogenesis? I don't see why. After all, several spermatogenesis genes, and even SRY, have already been dumped in other species with no ill effect.

Indeed, the book abounds with bold assertions hedged by "I can't prove it but...". Families that produce more boys than girls (the Sykes clan again, documented by dusty records from a village school) expose a superselfish Y chromosome. Newspaper accounts of female-only families are proof of toxic, Y-hating, superselfish mitochondria. Even the 'gay gene' turns out to be a mitochondrial plot.

I welcome speculation in popular-science books. Sharing with the public the leaps of imagination that make science exciting and creative might banish its image as gadget-driven and boffin-dominated. But speculation on speculation becomes tedious, and ultimately I feel that the central argument degenerates under its weight — like the Y chromosome itself. ■

Jennifer A. Marshall Graves is in the Research School of Biological Sciences, Australian National University, Canberra, ACT 2601, Australia.

Putting on the brakes

Henry Harris

It is a common assumption that somatic cells do not multiply unless they are stimulated to do so, and if they do not receive the necessary external or internal stimuli they are said to remain in the 'resting' phase. But this assumption is far from self-evident. It is entirely plausible to regard exponential multiplication, and not repose, as the cell's natural steady state. As a result of evolution, the cellular machinery for the conversion of nutrients into energy is geared towards cell multiplication. If this was not the case, it is difficult to see how natural selection, as it is currently envisaged, could operate. If, given an adequate nutrient supply and a clement physical environment, a population of cells does not multiply, those cells are not 'resting', they are repressed. The question then is not what induces cells to multiply, but what restrains them from doing so. Leaving aside conceptually unproblematic factors such as external toxicity, there appears to be only one process that represses cell multiplication under physiological conditions, and that is differentiation. Differentiation determines tissue specificity, and in doing so, it may suppress multiplication altogether, as it largely does in the central nervous system. Alternatively, it may permit multiplication to continue, but under severely restricted and regulated conditions, as in the intestine or the bone marrow. In the extreme case, it can even encompass the elimination of the cell nucleus or programmed cell death.

What, then, do we mean by oncogenes and tumour-suppressor genes? We surely do not mean that evolution generated specific genes to induce or suppress the growth of tumours.

In the case of tumour-suppressor genes the position is clear. The multiplication of tumour cells is suppressed by the same set of genes as those that suppress the multiplication of normal cells of the same type during the process of differentiation; tumours arise when these genes are impaired. In *Drosophila* the experimental evidence is conclusive. Tumorous growths in the developing larva are produced at specific sites and at specific times by the impairment of genes that have critical roles in the process of normal larval differentiation. So far, such precision has not been possible with mammalian cells, but the observations that have been made are entirely consistent with this conclusion. In hybrids formed by fusing a range of different malignant tumour cells with normal diploid fibroblasts, tumorigenicity is systematically suppressed when the composite cell retains the ability to execute the differentiation programme of the fibroblast. But when the ability to execute this programme is lost, tumorigenicity reappears.

A great deal of information has been accumulated about the *retinoblastoma* (*Rb*) gene (the first mammalian tumour-suppressor gene to be isolated), the protein that it encodes and the biochemical interactions in which this protein takes part. But why, when the *Rb* gene is eliminated, is a tumour of the retina by far the commonest malignancy formed? Elimination of the *Rb* gene does indeed increase the incidence of tumours at other sites, but at a level incomparably lower than the incidence of tumours of the retina. This finding indicates that the *Rb* gene is involved in some way in the mechanisms that determine tissue specificity, but the nature of this involvement awaits a thorough molecular analysis of differentiation in the mammalian retina. The same is true for tumours at other sites. Molecular biology has provided many interesting, but partial, insights into the modes of action of tumour-suppressor genes, but in no case that I am aware of has the causal nexus between differentiation and the restraint of cell multiplication been satisfactorily elucidated in molecular terms.

And what about oncogenes? It was originally assumed that 'resting' cells require stimulation in order to multiply, and that cells

Tumour suppression

Viewing cancer as a disease of cell differentiation rather than multiplication allows a redefinition of the role of oncogenes and tumour-suppressor genes.

in which multiplication is under strict physiological control require further stimulation if their multiplication is to be unbridled. That now seems very unlikely. If what has been said about the control of cell multiplication is true, then there is, in principle, only one way that oncogenes can influence cell multiplication under physiological conditions, and that is by impeding the operation of those genes that restrain this multiplication during the course of differentiation — the tumour-suppressor genes. Oncogenic mutations (and oncogenic viruses) release the brakes that tumour-suppressor genes apply.

It is often said that the growth of tumours is determined by the balance between the activity of oncogenes and that of tumour-suppressor genes. But the nature of this 'balance' is not at all clear. Is it really being proposed that the cell in some way titrates the cumulative effect of one set of genes against the cumulative effect of another? We know that a single tumour-suppressor gene may be involved, with varying degrees of efficacy, in the suppression of tumorigenesis in several different tissues. But we do not know how many genes are required to effect this suppression in any one case. It is possible that during the process of differentiation, many genes may be required acting combinatorially or sequentially or both. In the apparently simple notion of 'balance' there is thus ample room for complexity. But however complex the ultimate phenotype of a malignant cell might be, it would reduce confusion if it could be agreed that cancer, in the first instance, is not a disease of cell multiplication, but a disease of differentiation.

Henry Harris is at the Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK.

FURTHER READING

- Cavenee, W., Hastie, N. & Stanbridge, E. (eds) *Recessive Oncogenes and Tumour Suppression* (Cold Spring Harbor Lab. Press, New York, 1989).
- Harris, H. *The Cells of the Body* 211–247 (Cold Spring Harbor Lab. Press, New York, 1995).
- Mechler, B. M. in *The Legacy of Cell Fusion* (ed. Gordon, S.) 183–198 (Oxford Univ. Press, Oxford, 1994).
- Harris, H. *J. Cell Sci.* **79**, 83–94 (1985).
- Harris, H. *DNA Cell Biol.* **22**, 225–226 (2003).



Cancer: is unrestricted cell multiplication due to a defect in differentiation?

Protein surgery

Hans-Georg Rammensee

Studies of human tumours and the immune system have revealed that cutting and pasting of proteins can generate new peptide variants. This startling finding has implications for both proteomics and immunity.

Question: how many different peptides composed of nine amino acids can a cell produce from a protein made up of 500 amino acids? Answer: 482 — the first peptide consists of amino acids 1 to 9 of the protein; the second contains amino acids 2 to 10; and so on to the last peptide, which comprises amino acids 482 to 500.

At least, that's what we thought. But the implication of the paper on page 252 of this issue, by Hanada and colleagues¹, is that protein splicing allows for a much higher number. In other words, by slicing a protein into pieces, stitching different portions together, and then cutting out strings of nine sequential amino acids from the melded pieces, cells can manufacture entirely new sets of peptides from the original protein.

Hanada and colleagues' work centres on the use of peptides by the immune system. Every cell in the body is covered with peptides composed of eight to ten amino acids, glued to receptors known as major histocompatibility complex (MHC) class I molecules. The peptides represent every protein that is being made in the cell, and are crucial in allowing the immune system to detect and eliminate intruders. If, for instance, a virus has infiltrated a cell, then the evidence of its presence will be displayed on the cell surface. The immune system's 'killer' T cells will detect that tell-tale sign and take steps to destroy the infected cell.

Previously, Hanada and colleagues had cloned a human killer T cell that they discovered infiltrating a patient's kidney cancer². They found that the T cells recognized a peptide derived from a particular cellular protein, fibroblast growth factor-5 (FGF-5), that is overproduced in the tumour. In this case, the peptide was presented on a type of MHC class I molecule called HLA-A*03, which is known to display nine-amino-acid peptides with a tyrosine residue at position 3 and a lysine or arginine at position 9.

But what is the precise FGF-5-derived peptide that is recognized by the killer T cells — in other words, what is the T-cell 'epitope'? To find out, Hanada *et al.*¹ prepared target cells expressing truncated FGF-5 genes. The aim was to make it easier to identify the peptide by narrowing down the region of FGF-5 to search. They discovered that cells expressing a 60-amino-acid chunk of FGF-5, amino acids 161–220, were recognized by the T

cells. Knowing the requirements of the MHC molecule, it should then have been a matter of routine to identify the T-cell epitope. But the T cells recognized none of the possible strings of eight, nine or ten sequential amino acids from the 60-amino-acid fragment.

A clue to the mystery came from the authors' finding that (accidentally?) omitting several amino acids from within this 60-amino-acid fragment did not prevent T cells from recognizing their targets. But cells expressing fragments that were shortened at either end were not recognized. This serendipity prompted the authors to analyse a series of more extended deletions. The smallest construct that allowed T-cell recognition consisted of amino acids 172–176 and 199–220 of FGF-5 — suggesting that what the peptide recognized was actually a patchwork, containing bits from two separate parts of the protein. Indeed, Hanada *et al.* found that a peptide consisting of five residues from one end of this construct and four from the other was the T-cell epitope in question.

How could a cell make such a cut-and-paste peptide? The authors first wondered whether RNA splicing was the answer. This is a process in which, after a gene has been copied into messenger RNA but before that mRNA is translated into protein, segments of the RNA are excised and the remaining fragments are spliced together. But Hanada *et al.* judged this to be unlikely, as none of the known 'signatures' of RNA splicing are present in the gene sequence. Could the answer instead be sloppy translation, with the protein-making machinery skipping stretches of mRNA and starting again later? Experiments ruled this possibility out. So a post-translational mechanism seemed likely.

Could the explanation be protein splicing? This would involve the cells in cutting the protein into pieces and sticking them back together in a different order. Similar processes of protein surgery have been observed before, in single-celled organisms and some plants. Hanada and colleagues — two of whom are, coincidentally, from a surgery department — find that it also occurs here. They show that cells can use a synthetic 49-amino-acid fragment, in which

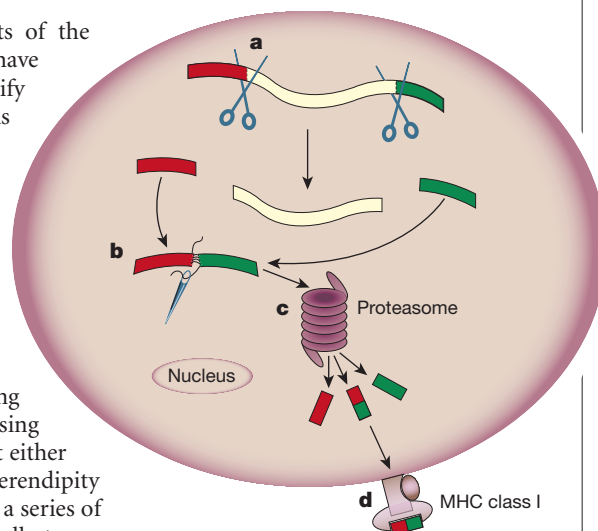


Figure 1 Protein splicing and the immune system. Killer T cells scrutinize short peptides displayed on MHC class I molecules on the surface of other cells. Hanada *et al.*¹ have discovered that the process of peptide generation from larger proteins can occur by protein splicing. **a**, A protein is cleaved by an (as yet unidentified) endopeptidase enzyme. **b**, Two of the pieces are sewn together (also by an unknown mechanism). **c**, The stitched-together intermediate is probably then sliced into shorter pieces by the proteasome. **d**, After further processing, nine-amino-acid peptides are glued to MHC class I molecules, to be displayed on the cell surface. Details such as cellular compartments, peptide transport and trimming 'exopeptidases' have been omitted for clarity.

the two ends of the T-cell epitope are separated by 40 amino acids, to construct the intact nine-amino-acid epitope. This could only happen if the cells were to cut the protein fragment into pieces, join two of the bits with a peptide bond, and funnel the new piece into the normal epitope-generating pathway (Fig. 1).

So far, two categories of natural protein splicing have been described. In single-celled organisms, controlled protein splicing is an important means of generating functional proteins. It is regulated by 'inteins' (intervening sequences) that catalyse their own excision out of a protein; the flanking fragments, or 'exteins', are then ligated^{3,4}. Inteins must have a particular structural domain if they are to catalyse this reaction, and the smallest

intain that can have such a domain consists of 134 amino acids^{5,6}. So it is unlikely that the events described by Hanada *et al.* are regulated in the same way. More similar is the process, seen in jack beans, of enzyme-mediated protein splicing — the ligation of polypeptide stretches to result in a functional protein⁷. In addition, protease enzymes, which generally slice up proteins, have been engineered to work in reverse⁸. Whatever the mechanism, this is, to my knowledge, the first time that protein splicing in human cells has been reported.

What are the implications of these findings? First, the potential number of different proteins and protein derivatives produced from our 30,000 genes increases enormously. Processes such as DNA recombination and RNA splicing were already known to increase this number; the discovery of protein splicing adds to the toolkit. Second, although we do not yet know how frequently protein splicing occurs, we have to consider the possibility that cells can produce non-continuous T-cell epitopes not only from tumour proteins but also from infectious agents and our own normal proteins — with implications for vaccine development and autoimmune diseases, as well as for cancer research.

Do mammalian cells use protein splicing to produce functional proteins or peptides that have any physiological role apart from the one leading to its discovery? We have no idea. The enzymes involved in protein splicing are also unknown. Those responsible for cutting might be conventional 'endopeptidases', including the cell's garbage-disposal unit, the proteasome. But what entity can join two protein fragments together? Can proteases work in reverse naturally, as well as being engineered⁸ to do so? Could the whole process of cutting and ligation happen inside the proteasome? Also, do sequence motifs dictate where within a protein splicing occurs? Here at least we might know the answer: given that the T-cell epitope described by Hanada *et al.*¹ is identical in every one of the kidney-tumour cells — and in other cells that are engineered to over-produce FGF-5 — this must be the case.

More questions asked than answered? Take it as an indication of an unexpected discovery. ■

Hans-Georg Rammensee is in the Department of Immunology, University of Tübingen, Auf der Morgenstelle 15, D-72076 Tübingen, Germany.
e-mail: rammensee@uni-tuebingen.de

1. Hanada, K.-i., Yewdell, J. W. & Young, J. C. *Nature* **427**, 252–256 (2004).
2. Hanada, K.-i. *et al.* *Cancer Res.* **61**, 5511–5516 (2001).
3. Noren, C. J., Wang, J. & Perler, F. B. *Angew. Chem. Int. Edn* **39**, 450–466 (2000).
4. Perler, F. B. *et al.* *Nucleic Acids Res.* **22**, 1125–1127 (1994).
5. Perler, F. B. *Nucleic Acids Res.* **30**, 383–384 (2002).
6. Perler, F. B. *Cell* **92**, 1–4 (1998).
7. Carrington, D. M., Auffret, A. & Hanks, D. E. *Nature* **313**, 64–67 (1985).
8. Jackson, D. Y. *et al.* *Science* **266**, 243–247 (1994).

Condensed-matter physics

Supersolid helium

John Beamish

Superfluids flow without resistance. It's hard to imagine, but quantum mechanically possible, that solids should do the same at low enough temperatures. Helium-4 might be the first known 'supersolid'.

At temperatures below 2.176 K, helium-4 enters a superfluid state and flows without friction. This 'perpetual motion' makes superfluidity — perhaps even more than its electronic counterpart, superconductivity — the most dramatic manifestation of quantum mechanics on a macroscopic scale. Despite its appeal, and despite many searches for superfluidity in other systems, it remains an uncommon phenomenon. From 1938, when superfluidity was discovered^{1,2}, helium-4 was the only known example until 1972 when the phenomenon was seen³, at much lower temperatures, in helium-3. That temperature difference between the two isotopes' behaviour reflects the intimate connection of superfluidity to Bose–Einstein condensation — the transition that occurs when 'bosons' collect in a single quantum mechanical state. Atoms of helium-4 are bosons, but helium-3 atoms are 'fermions' and must pair up before they can condense into a single state.

In 1995, advances in laser-cooling and magnetic-trapping techniques led to the achievement⁴ of Bose–Einstein condensation in rubidium vapour, adding to the list of superfluid systems. That list now includes other gases, such as spin-polarized hydrogen gas⁵, and, most recently, molecular gases of paired fermions^{6,7}. On page 225 of this issue, Kim and Chan⁸ claim the first observation of superfluid behaviour in a solid. A sample of solid helium-4, confined in the nanoscale pores of Vycor glass and rotated in a torsional oscillator, underwent a transition below about 175 mK that indicated the onset of 'supersolid' behaviour (Fig. 1). If it can be confirmed that superflow is occurring in the solid helium, this is a remarkable result indeed.

Despite their rarity, superfluids are fundamental to, for example, statistical mechanics and fluid dynamics, and they are a valuable test bed in fields as diverse as turbulence and cosmology. So it is not surprising that superfluidity has been sought in new systems. Solids, with their atoms localized on a periodic lattice, are certainly the most unexpected phase of matter in which to find superfluidity. However, the low atomic mass and the weak interatomic forces in solid helium make it very different from conventional solids. A quantum mechanical effect known as 'zero-point motion' dominates its properties, to the extent that it does not

freeze at all unless external pressure (of at least 25 bar) is applied. At higher pressures helium does crystallize, but zero-point motion remains important and produces a very compressible low-density solid. In such a 'quantum solid', defects such as vacancies on the atomic lattice are easily created and can be very mobile. They might even exist in finite concentrations at absolute zero. Although such zero-point vacancies have not been directly observed, they would be expected to condense into a coherent state at low temperatures. And because mass flow accompanies the movement of vacancies, such a state could exhibit superfluid flow^{9,10}.

Kim and Chan's results⁸ suggest supersolid behaviour in helium that is confined in the pores of Vycor. Vycor is a glass with a disordered network of nanoscale pores, and it has been widely used to study confinement effects in both liquid and solid helium. Confinement tends to suppress freezing, and substantially higher pressures are required to solidify helium in Vycor than when it is not confined. The authors suggest that Vycor's small pores should also enhance the concentration of vacancies in the solid helium, which might explain why they were successful when similar measurements with bulk samples of helium-4 did not detect any supersolid.

If this discovery of a supersolid is confirmed, it is a major advance for the field, because much of our understanding of superfluids is based on the translational invariance that is absent in a periodic solid. There are many open questions, for example the vacancy concentration required for supersolidity and its relationship to fundamental quantities such as the fraction of a sample that is superfluid and the critical velocities therein. There is also an interesting connection to recent work¹¹ on Bose-condensed gases confined in a periodic optical lattice of laser beams, which showed a transition from superfluid to insulator as the lattice potential was increased.

Kim and Chan's claim is sure to generate interest — and some controversy. Their torsional-oscillator technique is based on the classic experiment by Andronikashvili¹², in which a stack of oscillating disks immersed in liquid helium was used to determine how much of the helium was being viscously dragged by the disks (the normal fraction) and how much had decoupled from the disks (the superfluid fraction). Kim and Chan

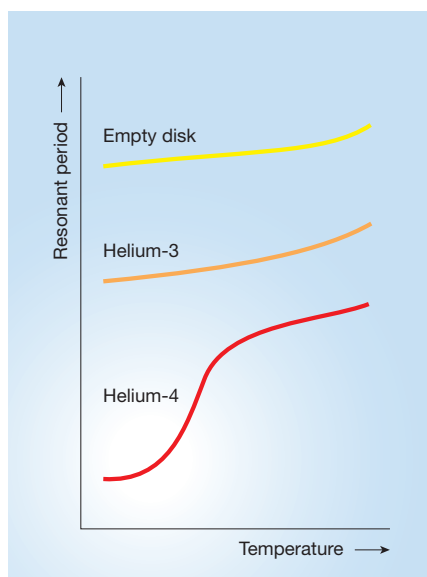


Figure 1 Super solid. Kim and Chan⁸ suspended a porous Vycor glass disk (15 mm in diameter and 4 mm thick) containing solid helium-4 in a torsional oscillator and monitored the disk's rotational inertia as the temperature decreased. Below about 175 mK they saw a sharp drop in the resonant period of oscillations, which is related to the disk's inertia. No such effect was seen for a Vycor disk empty of helium-4, or when the disk was filled with helium-3. Helium-4 was the first superfluid to be discovered; this might be the first evidence of supersolid behaviour.

have checked their measurements carefully to ensure that the decoupling they observe is real. They have also done the critical comparison with solid helium-3, for which they expect — and see — no decoupling.

Given the nearly atomic dimensions of the Vycor pores, it is hard to imagine that anything other than a superfluid or supersolid could move through them without dissipation. What is less certain is the exact

nature of the superflow that Kim and Chan have detected. Although their measurements were done at a pressure of 62 bar, well above the roughly 40 bar needed to solidify helium-4 in Vycor, it is possible that a disordered, liquid-like layer of helium could have remained near the pore walls. Even if such a layer is responsible for the superflow, its high density and the very small critical velocities observed by the authors imply that it must still be different from the superfluidity previously seen for thin films of liquid helium-4 in Vycor and other porous media.

The possible discovery of a new phase of matter, a supersolid, is exciting. The quantum-fluids and quantum-solids community can be expected to test the authors' claims thoroughly, particularly by searching for the persistent currents that are the 'gold standard' test of superfluidity. It will be fascinating to see how robust the phenomenon is; can it be seen in other porous media or even in bulk helium? There are enough questions to be answered about the nature and properties of supersolid helium to keep both experimentalists and theorists busy for a long time. ■

John Beamish is in the Department of Physics, University of Alberta, Edmonton, Alberta T6G 2J1, Canada.

e-mail: beamish@phys.ualberta.ca

1. Kapitza, P. *Nature* **141**, 74 (1938).
2. Allen, J. F. & Misener, A. D. *Nature* **141**, 75 (1938).
3. Osheroff, D. D., Richardson, R. C. & Lee, D. M. *Phys. Rev. Lett.* **28**, 885–888 (1972).
4. Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. *Science* **269**, 198–201 (1995).
5. Fried, D. G. et al. *Phys. Rev. Lett.* **81**, 3811–3814 (1998).
6. Greiner, M., Regal, C. A. & Jin, D. S. *Nature* **426**, 537–540 (2003).
7. Jochim, S. et al. *Science* **302**, 2101–2103 (2003).
8. Kim, E. & Chan, M. H. W. *Nature* **427**, 225–227 (2004).
9. Andreev, A. F. & Lifshitz, I. M. *Sov. Phys. JETP* **29**, 1107–1113 (1969).
10. Chester, G. V. *Phys. Rev. A* **2**, 256–258 (1970).
11. Greiner, M., Mandel, O., Esslinger, T., Hansch, T. W. & Bloch, I. *Nature* **415**, 39–44 (2002).
12. Andronikashvili, E. L. *Zh. Eksp. Teor. Fiz.* **16**, 780 (1946).

Palaeontology

Lost children of the Cambrian

Graham E. Budd

The initial flowering of animal life on Earth occurred during the Cambrian, some 540–490 million years ago. Fossil embryos from that time can provide clues about the origins of the major animal groups.

When ancient fossil 'trilobite embryos' were reported¹ from China in 1994, the reaction was largely sceptical. After all, biologists had been lamenting (or crowing) for years that fossil embryos could never be found. This bright certainty became mottled with doubts, however, as increasingly convincing material began to appear², the oldest and most controversial being some 600 million years old³.

These fossils raise several questions, to

say the least. First, how could they possibly be preserved? Second, why are they concentrated in a period (600–500 million years ago) that is already unfairly overstocked with exceptionally preserved fossils, such as those of the Burgess Shale in the Canadian Rockies? Third, do they tell us anything about animal evolution?

The preservation of these fossils is slowly being recognized as the product of the unusual geochemistry of the transitional



100 YEARS AGO

Dr. Nordenskjöld and the members of his South Polar Expedition arrived at Hamburg on January 6. The unexpectedly early return from the South Polar regions of this expedition has, the *Times* states, enabled Dr. Jean Charcot to recast the plans of the French expedition on board the *Français*. He now proposes to explore the west coast of Graham Land and to carry out a very exhaustive scientific investigation of that region... It is Dr. Charcot's definite intention to return at the end of the season of 1904–5. The *Français*, indeed, is only provisioned for two years, and Dr. Charcot states that if the expedition does not return in the early months of 1905, it must be concluded that they have been involuntarily detained, and a relief vessel must be dispatched to their assistance.

From *Nature* 14 January 1904.

50 YEARS AGO

On December 9, the P.E.N. (Poets, Playwrights, Editors, Essayists and Novelists) Club held an informal discussion on poetry and science... At first sight it might appear that these branches of culture had little to do with each other. But Prof. Dingle gave instances of the antagonism between poets and scientists, and pointed out that the overt attacks arising from this antagonism in the past were made by the poets. This, he said, is understandable; at that time it was thought that there was a real external world, the truth about which was being increasingly found out by the scientists. To poets, this world seemed flat and distasteful, but at the back of their minds the uncomfortable thought grew that any alternative was mere illusion. In Prof. Dingle's opinion, however, there is no need for the poet to harbour such resentment... In fact, science is the organized description of the relations between experiences; poetry, the expression of the experiences themselves. This description of the limits of science, abnegatory as it might seem, did not dispel the injured suspicion of the poets present. Stephen Spender... said that the psychologist may attribute the cause of conscience to infantile experiences; he himself might attribute it to God; what claim had science to the unique truth of the matter? Not only are the findings of science uncomfortable in detail; they are also so complicated that no one man can understand them, and increase of understanding brings with it disorientation and despair.

From *Nature* 16 January 1954.

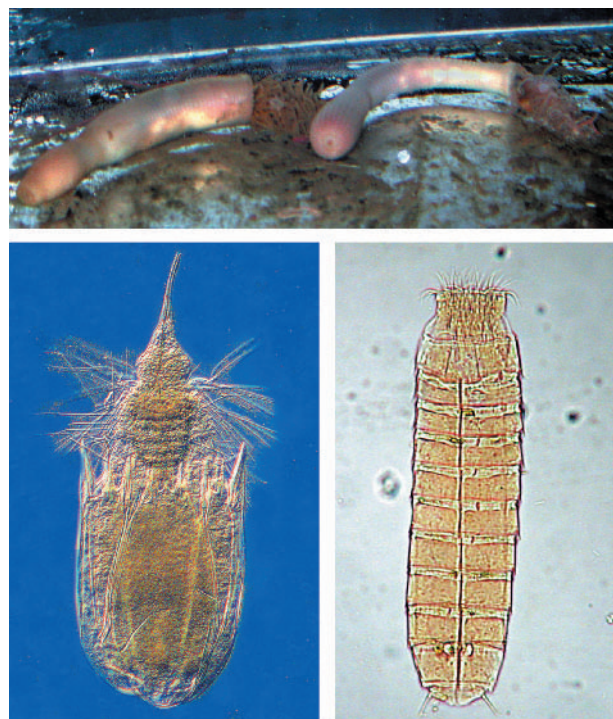


Figure 1 Adult representatives of the three living scalidophorans. Top, two specimens of the priapulid *Priapulid caudatus* (length about 10 cm). Bottom left, the loriciferan *Nanaloricus mysticus* (0.3 mm). Bottom right, the kinorhynch *Pycnophyes kielensis* (0.7 mm). The Cambrian embryo *Markuelia*, the subject of Dong and colleagues' analysis⁵, seems to be more primitive than any of these very diverse worms. (Images of *Nanaloricus* by R. M. Kristensen, and of *Priapulid* and *Pycnophyes* by G. E. Budd.)

time between the Cambrian period and the preceding Proterozoic era. But their evolutionary significance has largely been moot, partly because of the problems of relating the embryos to known groups of animals. All the convincing examples so far have been of apparently 'direct-developing' animals — animals that, unlike most living marine invertebrates, do not have a free-swimming planktonic stage. This apparent lack in the Cambrian suggested that direct development was the primitive state⁴ and that indirect development was something that evolved later, in contradiction of many fashionable theories of animal evolution.

On page 237 of this issue, Dong *et al.*⁵ present new material of a fossil — *Markuelia* — that was recognized as an embryo a few years ago², but in only enough detail to tantalize palaeontologists about its adult affiliations. The inevitable speculation about its adult form brought in members of various fossil groups^{2,4}, including those with a claim to being the last common ancestor of annelids, molluscs and brachiopods. Dong *et al.* reconstruct the embryonic *Markuelia* as an annulated worm, with a mouth at one end that is surrounded by a series of spines or scalids.

Their new data suggest comparison with a completely different group of animals whose time for world attention has at last arrived. These are the scalidophorans⁶, a collection of three gloriously obscure marine worms: priapulids, loriciferans and kinorhynchs (Fig. 1). Molecules and morphology ally these worms with that model of model organisms, the nematode *Caenorhabditis elegans*, although all these living animals differ greatly one from the other. The reason why this motley assortment has suddenly become rather interesting is that molecular

data place them and the nematodes, not as some sort of unremarkable scion of early animals, but as the closest relatives⁷ to the most important group of all, the arthropods — which includes insects and crustaceans — whose species are numbered in their millions.

Basic embryological data are frustratingly patchy for the living scalidophorans, so the bizarre possibility of their embryology being described better from the fossil record than from the living forms cannot be ruled out — a true reversal of fortune. Dong *et al.*⁵ analyse the relationships of the fossils by using cladistics, the methodology of choice for most serious systematists. They show that *Markuelia* does not fall into any of the living major groupings within the scalidophorans, but instead lies in their 'stem-group'⁸: that is, it is related to them all but is more primitive than any living form (Fig. 2).

Thus, Dong *et al.* claim (and convincingly) that *Markuelia* takes us back to a pattern of embryonic development more 'basal' than any of the living groups. Understanding this

pattern could, for example, provide clues about whether the arthropods inherited major features such as segments, blood vascular systems and dorsal brains from a very deep ancestor of animals or evolved them anew. All these features are apparently missing from the living scalidophorans and nematodes. But did they never have them, or have they been lost? Here, *Markuelia* already offers one hint of potentially profound significance. The surface annulation of the worm penetrates into the body cavity in the form of septa. Is this a suggestion of segmentation arriving in — or leaving — these animals?

How do these evolutionary questions tie in to the preservational ones? Perhaps the reason why only direct developers are found in the Cambrian is that they tend to have a robust egg shell (or yolk eggs²), which unlike the delicate free-swimming forms could act as a template for early mineralization. If so, then the lack of Cambrian indirect developers is apparent rather than real. Further, *Markuelia* is now revealed to sit in the right taxonomic place to be a direct developer, just like all living scalidophorans and nematodes. As Dong *et al.* rightly stress, the palaeontological data here corroborate and extend that from living animals.

The miraculous preservation of these embryos comes with an annoying drawback: it seems to work only with very small fossils. As a result, the embryos known are largely orphans: we have little clue about their adult parents. Still, it is likely that no profound metamorphosis took place in *Markuelia*, and seasoned Cambrian-watchers will no doubt be searching for suitable candidate adults. Given the enormous size increase that takes place after hatching in living priapulids, it is likely that the adult *Markuelia* was quite big: even as an embryo it is larger than most living adult scalidophorans. In any case, the prospect of having complete life cycles described from the Cambrian² no longer looks fanciful; and with that, the prospect of making direct comparisons of ancient and living life cycles comes into view. It looks as though the study of 'evolution of development', which has hitherto relied greatly

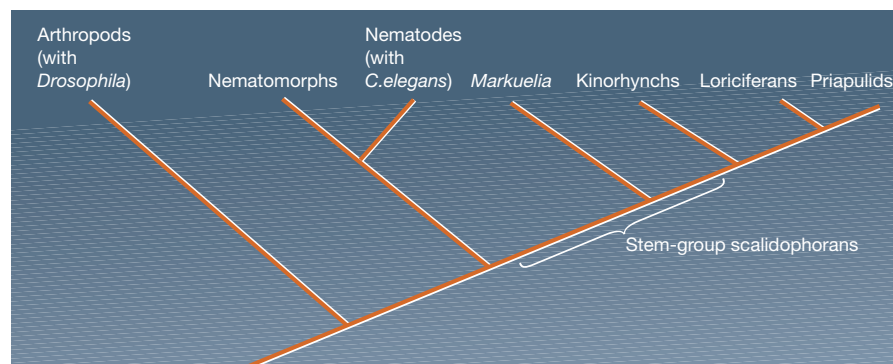


Figure 2 The place of *Markuelia* in evolution, as inferred from the new work⁵. The positions of the two major model organisms, *C. elegans* and *Drosophila*, are also indicated. Nematomorphs are a group of parasitic worms closely related to the nematodes.

on comparative gene expression data, is also going to have to start grappling with the message of the lost children of the Cambrian too.

Graham E. Budd is in the Department of Earth Sciences, Palaeobiology, Norbyvägen 22, University of Uppsala, Uppsala 752 36, Sweden. e-mail: graham.budd@pal.uu.se

1. Zhang, X.-G. & Pratt, B. R. *Science* **266**, 637–639 (1994).
2. Bengtson, S. & Yue, Z. *Science* **277**, 1645–1648 (1997).
3. Xiao, S., Yang, Z. & Knoll, A. H. *Nature* **391**, 553–558 (1998).
4. Conway Morris, S. *BioEssays* **20**, 676–682 (1998).
5. Dong, X.-p., Donoghue, P. C. J., Cheng, H. & Liu, J.-b. *Nature* **427**, 237–240 (2004).
6. Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla* 2nd edn (Oxford Univ. Press, 2001).
7. Aguinaldo, A. M. A. *et al.* *Nature* **387**, 489–493 (1997).
8. Budd, G. *Nature* **412**, 487 (2001).

Earth science

Keeping score on the core

Erik Hauri

Getting to the bottom of events at the boundary between Earth's core and mantle is fiendishly difficult. The latest analysis invokes evidence from an isotope of tungsten to conclude that the two do not interact.

The report by Scherstén and colleagues (page 234 of this issue¹) bears on one of the liveliest debates in Earth science — the extent to which Earth's core exchanges matter with the mantle. This topic has been addressed by many groups, and from a number of observational, experimental and theoretical directions². In particular, however, many analyses have centred on the geochemistry of highly metallic elements in mantle rocks, with recent efforts focusing on volcanic hotspots such as Hawaii. Hotspots form above plumes of material, rising up through the mantle, that may (or may not) originate from the core–mantle boundary at the base of the mantle. Over the past few years, persuasive arguments in favour of core–mantle exchange have been made on the basis of osmium isotopes in hotspot lavas, especially the small anomalies in ¹⁸⁶Os that may have been derived from the decay of

an isotope of platinum (¹⁹⁰Pt) in the Earth's outer core^{3,4}.

Scherstén *et al.*¹ take a new approach to the issue of core–mantle exchange, by looking for anomalies in the isotopes of tungsten (W) in mantle-derived lavas. Tungsten is a highly siderophile — 'iron-loving' — element thought to be present in large quantities in the Earth's core. Its parent isotope ¹⁸²Hf decays radioactively to ¹⁸²W with a geologically rapid half-life of about 9 million years. As a result, all of the ¹⁸²W derived from ¹⁸²Hf decay was formed within the first 60 million years of Earth's history at a time that encompasses the formation of the Earth's core^{5–8}. Searching for anomalies in ¹⁸²W is thus a powerful approach to studying the flux of material from the core to the mantle, because all of the action in the hafnium–tungsten isotope system occurred in the first 60 million years of Earth's history.

Subsequent processes (such as mantle convection, formation of the continental crust and plate subduction) serve only to mix the mantle and thus reduce any primordial differences in ¹⁸²W that may have existed between different parts of the mantle.

It is widely accepted that the Earth formed from chondritic (stony) meteorites, material in the early Solar System that accreted into increasingly large bodies. High-precision data on ¹⁸²W in chondritic meteorites^{5,6} have shown that the Earth's mantle is comparatively enriched in ¹⁸²W (by 2 parts in 10,000, or $\epsilon W = +2$). This means that the Earth's iron core separated from the silicate mantle within the first 30 million years of accretion, while ¹⁸²Hf was still 'alive' and producing ¹⁸²W by radioactive decay. With most of the tungsten partitioned into the core and most of the hafnium remaining in the mantle, the Earth's core must be depleted in ¹⁸²W (though, obviously, this has not been directly measured), whereas the mantle is clearly enriched compared to chondritic meteorites — the $\epsilon W = +2$ enrichment of the Earth's mantle in ¹⁸²W has been measured directly and independently by various groups^{5,6}.

Thus, variations in ¹⁸²W in mantle-derived rocks should faithfully record interaction between the core and mantle — if it takes place. The idea is that if small amounts of material from the Earth's core are added to the mantle, then such mantle should show depletions in ¹⁸²W (see Fig. 1a, overleaf, which depicts the preferred model of Scherstén *et al.*). Scherstén and colleagues' measurements appear to be of very high quality, and the homogeneity of these plume-related samples suggests that the modern mantle is homogeneous in tungsten isotopes. In particular, the tungsten measurements were

Astronomy

Star maker

Star formation can occur wherever interstellar gas becomes compressed and undergoes gravitational collapse. That compression might be caused by a supernova explosion, or the collision of galaxies. At last week's meeting of the American Astronomical Society in Atlanta, Georgia, Richard Rees and Kyle Cudworth presented the first evidence for star formation triggered by a new mechanism — a globular cluster of stars that punched through the disk of the Milky Way five million years ago.

The globular cluster NGC 6397 (pictured), 12 billion years old and containing a few hundred thousand stars, currently sits about 1,500 light

years below the disk of the Milky Way. Its motion has been tracked by various telescopes since 1893, so Rees and Cudworth were able to extrapolate its path back, right through the plane of our Galaxy. And at the point at which the cluster crossed the disk five million years ago lies NGC 6231, a young cluster of stars estimated to have formed less than five million years ago.

The implication is that the passage of NGC 6397 through the Milky Way disk compressed the gas in that region and triggered the birth of NGC 6231 — a mechanism for star formation that was suggested in a little-noticed 1996 paper



(J. F. Wallin *et al.* *Astrophys. J.* **459**, 555–557; 1996). There it was noted that a globular cluster passes through the disk of the Milky Way roughly every million years. So it

seems that the star-making talent of globular clusters should be factored into our picture of the evolution of this, and other, galaxies.

Alison Wright

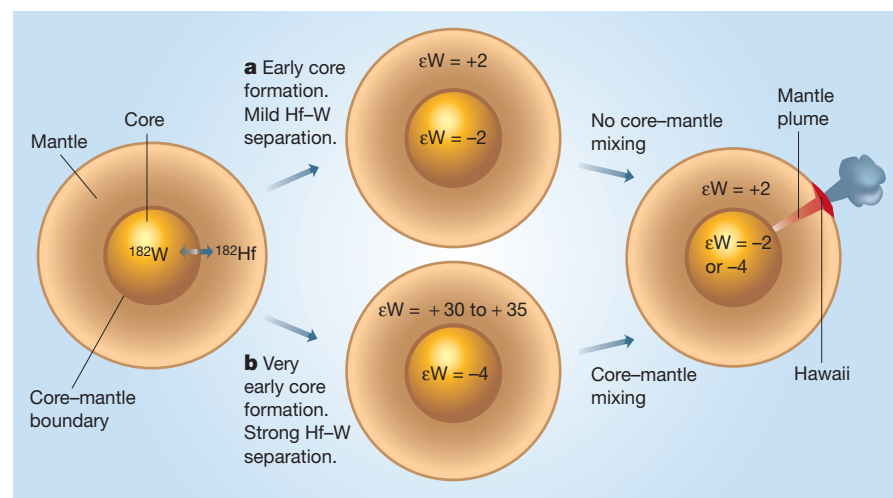


Figure 1 Alternative interpretations of core-mantle interaction as surmised from the tungsten isotope ^{182}W . Formation of the Earth's core (left) results in separation of tungsten into the core along with iron, whereas its parent hafnium (^{182}Hf) remains in the mantle. **a**, In the preferred model of Scherstén *et al.*¹, the core forms early (some 30 million years after accretion of the Earth), with only mild separation of hafnium and tungsten. The core subsequently remains isolated from the mantle. **b**, An alternative model postulates very early core formation (say, within 5 million years), stronger hafnium-tungsten separation, and mixing at the core-mantle boundary which results in transfer of core material into the mantle. Both models could produce a modern mantle with a tungsten isotope anomaly of $\epsilon\text{W} = +2$, as assessed from rocks, such as those on Hawaii, that are produced by an ascending mantle plume.

made on exactly the same Hawaiian rock samples that displayed enrichments in ^{186}Os attributed to a contribution from the core, and there is no correlation between the ^{186}Os and ^{182}W isotopes.

The authors thus draw the most obvious conclusion — that lack of variation in the tungsten isotopes, and absence of a correlation with ^{186}Os , must mean that there is no contribution from the core in these Hawaiian hotspot samples. Nor, they believe, is there any such contribution in other, South African samples derived from the deep mantle (kimberlites) that they analysed. This is the simplest and most straightforward explanation of the data. The authors outline at least one way — recycling of oceanic plates in the convecting mantle — in which the ^{186}Os anomalies can be explained in the absence of ^{182}W anomalies.

This conclusion is an important one — it implies that the Earth's core has remained perfectly isolated from the mantle since it formed, and that the only thing coming out of the core is heat.

However, there remain tantalizing clues that the story may not be so simple. Mantle concentrations of highly siderophile elements have long been known to be much higher than those predicted from the behaviour of such elements in high-temperature core-formation experiments². Most of these experiments were conducted at much lower pressures than those pertaining at the core-mantle boundary. So the mantle enrichment in highly siderophile elements might be merely the result of a higher pressure of core formation, and their different behaviour under such conditions.

Yet there is another possible model (Fig. 1b). It is conceivable that the differences in the ^{182}W abundances in the core and mantle are even larger than those modelled by Scherstén and colleagues. If the Earth's core formed and removed tungsten from the mantle much earlier than previously

thought (say, within the first 5 million years of Earth's formation), the core could have a depleted ^{182}W abundance as low as $\epsilon\text{W} = -4$, similar to that of some iron meteorites that are themselves the ancient cores of fragmented protoplanets^{7,8}. If so, then ^{182}Hf decay in the mantle over the next 60 million years will have produced an enormous anomaly in ^{182}W (perhaps as high as $\epsilon\text{W} = +30$ to $+35$). In this case, the only way to produce a modern mantle with a small ^{182}W anomaly ($\epsilon\text{W} = +2$) is to postulate very early core formation on the Earth, and then later transfer of some material from the core to the mantle over the next 4.5 billion years of Earth's history⁹. Still, Scherstén *et al.* do outline a convincing case for isolation of the core from the remainder of the Earth, and their work should provide fertile ground for improved ^{182}W studies in all manner of rocks and meteorites.

Erik Hauri is in the Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road NW, Washington DC 20015, USA.

e-mail: hauri@dtm.ciw.edu

1. Scherstén, A., Elliott, T., Hawkesworth, C. & Norman, M. *Nature* **427**, 234–237 (2004).
2. Walker, D. *Geochim. Cosmochim. Acta* **64**, 2897–2911 (2000).
3. Brandon, A. D. *et al. Science* **280**, 1570–1573 (1998).
4. Brandon, A. D. *et al. Earth Planet. Sci. Lett.* **174**, 25–42 (1999).
5. Yin, Q. *et al. Nature* **418**, 949–952 (2002).
6. Kleine, T. *et al. Nature* **418**, 952–955 (2002).
7. Harper, C. L. & Jacobsen, S. B. *Geochim. Cosmochim. Acta* **60**, 1131–1153 (1996).
8. Lee, D.-C. & Halliday, A. N. *Science* **274**, 1876–1879 (1996).
9. Hauri, E. H., Carlson, R. W. & Bauer, J. *Lunar Planet. Sci. XXXI*, 1812 (2000).

Evolutionary biology

Our relative genetics

David Penny

Data on the chimpanzee genome help in detecting differential selection on individual genes, and in judging whether normal microevolutionary processes are sufficient to account for human origins.

Icreasingly accurate versions of the human genome sequence are being produced. But to find out what biologically makes us human we also need the sequenced genome of our nearest relative, the chimpanzee, to see whether there is anything in our genetic constitution that could not have arisen by well-understood genetic processes. Last month, a selected version of the chimpanzee genome was published by Clark *et al.* in *Science*¹. The group sequenced about 200,000 protein-coding regions, and combined them with human sequences to give over 20,000 chimp-human gene alignments. The mouse genome was then used as an 'outgroup', standard practice in this form of analysis, to leave 7,645 chimp and human genes for comparative analysis.

The fundamental issue here is Darwin's

bold claim that "numerous, successive, slight modifications" are sufficient for all of evolution (Fig. 1). This can be paraphrased, in later terms, as "microevolution is sufficient to explain macroevolution"². The historical context is that evolutionary biology can be divided into two phases³: first, the acceptance in the 1860s that evolution (macroevolution) had indeed occurred; second, the realization in the mid-1900s that the processes of microevolution (natural selection working through genetics) were necessary for evolution to occur.

Over the past 30 years, with the rise of molecular biology, the search has been on for support for Darwin's claim and to demonstrate the sufficiency², not just the necessity, of slight modifications in explaining macroevolution. The changes seen within populations and closely related species are by

definition 'slight modifications', and at the genetic level they take many forms — point mutations, small insertions, deletions and duplications of genes, chromosomal inversions and fusions, activation of transposable genetic elements of many kinds, and so on.

To test rigorously for the sufficiency of microevolution in human evolution, the full sequence of the chimpanzee genome will be necessary. But Clark *et al.*¹ make a good start on the task with their tests for the effects of natural selection on those genes that encode proteins. Use of the mouse genome as an outgroup allows estimates of the number of synonymous (silent) mutations and non-synonymous (replacement) mutations. The ratio of the two permits the potential identification of genes that have been under positive selection in humans as opposed to chimpanzees, and vice versa. Not surprisingly, selective changes occur in both the human and chimpanzee lineages (our common ancestor was neither chimp nor human).

One of Clark and colleagues' findings is that human enzymes for amino-acid breakdown (catabolism) have been under positive selection. This is concordant with the generally high proportion of meat (and thus protein) in the human diet, at least in comparison with the more herbivorous chimpanzee and gorilla. The increased capacity to break down amino acids is not surprising in another respect. For example, failure to catabolize phenylalanine has several adverse effects, including brain damage. Overall, the finding lends support to theories⁴ that an increased proportion of meat in the diet of early humans was important for an increase in brain size. Regardless of that, there could also be ethical implications. If early humans ate meat 'naturally', then for example being vegetarian could be considered a personal choice rather than a universal ethical decision. But all that can be claimed here is that scientific knowledge will be necessary, even if not sufficient, for solving such ethical questions.

Complex interactions are evidently occurring between the various genes involved in human olfaction⁵. The extensive results of Clark *et al.*¹ provide a wider picture of the consequences, which range from genes being under positive selection to those that have lost function (pseudogenes) — or those that may be in the process of losing function. These results illustrate how genome-wide information will stimulate new experiments, both at the level of gene expression and with the aim of making physiological comparisons. What, for instance, is the comparative sensitivity of humans and chimps to a range of olfactory stimuli? Do humans have an improved receptivity to odours from the increased proportion of meat and/or cooked foods in our diet? Such tests will allow us to see how genetic differences manifest themselves at the level of the organism, and we



Figure 1 The chimpanzee, our closest relative and a genomic test for Darwin's bold claim: "If it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down. But I can find no such case." (C. Darwin, Ch. VI, "Difficulties of the theory", *On the Origin of Species* 6th edn; 1872.)

can expect a burst of experiments to that end.

For the past 30 years, attempts to explain the differences between humans and chimps have centred on gene regulation⁶, rather than on the genes themselves. Those regulatory changes might be difficult to find in genome analyses, even though microarray experiments can now estimate changes in levels of gene expression⁷. But it is not clear whether genes with conserved, or with changed, levels of expression are the more important. Early in molecular-evolution studies it was assumed that differences in amino-acid sequences were the crucial ones. The neutral theory⁸ challenged this view and showed that the conserved regions of a gene were significant. The relative importance of conservation or change in gene expression remains to be evaluated. At the least, however, finding positive selection^{1,5,9,10} in many genes involved in metabolism and in physiology broadens our ideas on the role of selection in human and chimpanzee evolution.

A further example of what will be possible with full human and chimp genomes is accurate estimates of rates and variability of all the possible microevolutionary genetic changes. It is simple to estimate the number of mutations per year or per generation⁹. But good estimates of the date of divergence between chimpanzees and humans are essential in evaluating microevolutionary change, and the result depends on whether calibration points are used from within¹⁰, or outside¹¹, the primates as a group. But using a point (nucleotide) mutation rate of about 10^{-9} bases per year¹⁰, for a genome size of about 3×10^9 bases, still gives an average of well over one mutation per generation.

Given such information, a long-term issue might emerge: is the trend towards exact replacement rates of human reproduction (two offspring per female) genetically sustainable? If the proportion of deleterious and slightly deleterious mutations is significant, then exact replacement reproductive

rates might lead to eventual genetic decline¹². There is no short-term concern here: the planet's ecological sustainability is a much more immediate worry. But it is an indication of what questions might arise with firmer knowledge of our genetic evolution.

The results of Clark *et al.*¹, then, provide plenty of food for thought. However, there is much more to come: the entire chimpanzee genome is being sequenced by a publicly funded consortium¹³, and some data are already available through GenBank. The full sequence will be available later this year, and further comparative analyses should lead to a definite answer as to whether there is anything in the human genome that is not accounted for by the normal microevolutionary processes. Is there a genetic continuum between us and our ancestors and the great apes? If there is, then we can say that these processes are genetically sufficient to fully account for human uniqueness — and that would be my candidate for the top scientific problem solved in the first decade of the new millennium. ■

David Penny is at the Allan Wilson Center for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand.
e-mail: d.penny@massey.ac.nz

- Clark, A. G. *et al.* *Science* **302**, 1960–1963 (2003).
- Riddiford, D. E. & Penny, D. in *Evolutionary Theory: Paths into the Future* (ed. Pollard, J. W.) 1–38 (Wiley, New York, 1984).
- Mayr, E. *One Long Argument: Charles Darwin and the Genesis of Modern Evolutionary Thought* (Harvard Univ. Press, Cambridge, MA, 1991).
- Aiello, L. C. & Wells, J. C. K. *Annu. Rev. Anthropol.* **31**, 323–338 (2002).
- Gilad, Y., Bustamante, C. D., Lander, D. & Pääbo, S. *Am. J. Hum. Genet.* **73**, 489–501 (2003).
- King, M. C. & Wilson, A. C. *Science* **188**, 107–116 (1975).
- Enard, W. *et al.* *Science* **296**, 340–343 (2002).
- Kimura, M. & Ohta, T. *Proc. Natl Acad. Sci. USA* **71**, 2848–2852 (1974).
- Carroll, S. B. *Nature* **422**, 849–857 (2003).
- Wildman, D. E., Uddin, M., Liu, G., Grossman, L. I. & Goodman, M. *Proc. Natl Acad. Sci. USA* **100**, 7181–7188 (2003).
- Arnason, U., Gullberg, A., Schweizer Burguete, A. & Janke, A. *Heredity* **133**, 217–228 (2000).
- Lynch, M. *et al.* *Evolution* **53**, 645–663 (1999).
- www.nhgri.nih.gov/11509418

Molecular biology

Breaking up is hard to do

Science **303**, 243–246 (2004)

Cells protect the integrity of their genome using a process called recombination. This process also occurs during meiosis — when sex cells are being formed — and is essential for the production of genetic diversity.

During recombination, chromosomes become joined and swap segments of genetic information; afterwards, they must separate. Bacterial enzymes that aid chromosome separation are known, but the identities of the mammalian proteins concerned are uncertain. Yilun Liu *et al.* report that a molecule called Rad51C may be involved. When the protein is absent from cell extracts, DNA separation is reduced. When Rad51C is added back, chromosomes can break apart.

The study suggests that Rad51C — or a protein associated with it — acts as a DNA-separating enzyme in mammalian cells. Its discovery should help researchers understand how recombination is completed within nuclei.

Helen R. Pilcher

Neuroscience

Neurons get dressed

Phys. Rev. Lett. **91**, 268101 (2003)

The network of neurons in the brain is supported by a mass of glial cells, whose functions include supplying the neurons with nutrients and removing dead tissue. A particular type of glia, called astrocytes, can also bind glutamate and thereby affect the communication between neurons. Yet most models of the electrical responses of neurons consider only 'naked' neurons, ignoring any interaction with the surrounding glia.

Suhita Nadkarni and Peter Jung have developed a model of electrical activity in the brain that includes such interactions — these neurons are fully 'dressed'. The model predicts that when a dressed neuron is

exposed to a small electrical stimulus, its action potential is more likely to spontaneously oscillate than previous estimates have suggested. Oscillations of the neuron's action potential are associated with the seizures that are common in epilepsy.

Astrocytes of people with temporal lobe epilepsy tend to have extra glutamate-binding receptors. Nadkarni and Jung suggest that the extra receptors make spontaneous oscillations among neurons more likely, supporting the idea that overexpression of glutamate receptors could cause epilepsy.

Mark Peplow

Martian geology

Ice sculpture?

Geophys. Res. Lett. doi:10.1029/2003GL018575 (2003)

The dry valleys of Mars remain as mysterious as ever. The notion of a warm, wet early Mars seems to be undermined by simulations that suggest that an atmosphere of carbon dioxide and water vapour might never have been thick enough to raise global temperatures to near water's freezing point. In other words, it might have snowed on Mars, but it was too cold to rain.

Michael H. Carr and James W. Head III propose that the large amounts of surface water required to carve out channels hundreds of kilometres long might have come from thick glaciers and ice sheets that melted at their base. The internal heat of the young planet, although more intense than it is today, is not likely by itself to have been sufficient to melt a veneer of ice at the planet's surface. But the melting point of ice is pressure-dependent: ice squeezed at the bottom of thick ice sheets melts more readily. If, as seems possible, Mars once had thick ice sheets extending to low latitudes, Carr and Head calculate that they could have undergone extensive basal melting. Ground temperatures of just 230 K could have been sufficient to induce the melting of snowpack and ice 50 m below the ice-sheet surface.

Phillip Ball

Astrophysics

A double-pulsar binary

Science doi:10.1126/science.1094645 (2004)

Pulsars offer a unique laboratory for studying relativistic gravity and plasma physics. So the discovery by A. G. Lyne and colleagues of a double-pulsar system — the first ever seen — will delight physicists of all flavours.

Late last year, the same group reported the discovery of a binary system comprising a pulsar and a neutron star (*Nature* **426**, 531–533; 2003). Now they say that the neutron star is in fact a second pulsar, whose radiation is severely moderated by its larger sibling.

Analysis of the pulse frequencies and of the pulsars' elliptical orbits has already yielded measurements of the masses of the two bodies with a degree of precision that would not have been possible unless both stars were pulsars. And the orientation of the two pulsars means that the radiation from one shines through the magnetized volume surrounding the other, giving an unprecedented opportunity to probe the region immediately around each pulsar.

Other observations of the system fit well with predictions from Einstein's theory of relativistic gravity. Furthermore, Lyne *et al.* believe that their study already validates the suggested sequence for the evolution of such a system: a companion star spins up the pulsar, before becoming a pulsar itself in a supernova explosion.

Mark Peplow

Neurodegenerative diseases

Plaque connection

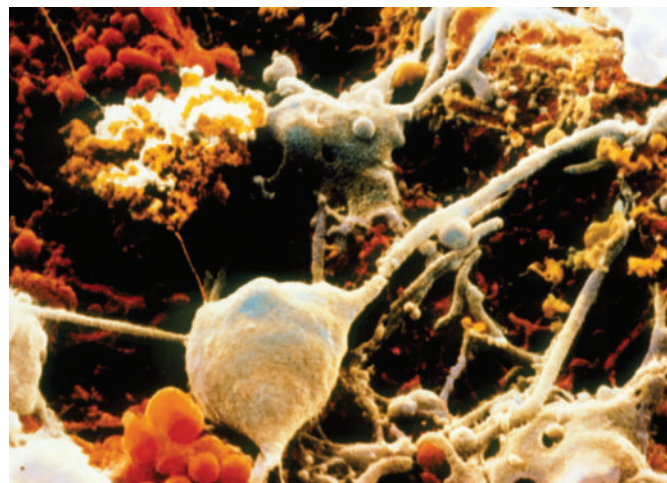
Neuron **41**, 27–33 (2004)

Amyloid plaques are a hallmark of Alzheimer's disease: these protein clumps, which are composed of the β -amyloid peptide, build up around brain cells. Yet the links between β -amyloid and dementia are unclear.

Masuo Ohno *et al.* now provide genetic evidence that soluble β -amyloid may trigger Alzheimer's disease. The team studied a strain of dementia-prone mice. Normally, these rodents have high levels of β -amyloid peptides and show cognitive impairment as they age — they become less able to remember cage mates and learned locations. But Ohno *et al.* find that when a key enzyme, called BACE1, is removed, the mice stay healthy and dementia free. BACE1 is needed for amyloid production; without it, amyloid levels remain low and nerve cells stay healthy.

The study suggests that elevated β -amyloid levels cause the cognitive decline associated with Alzheimer's disease. The authors hope that BACE1-inhibiting drugs will stave off or even reverse dementia in humans.

Helen R. Pilcher



Dressed for the occasion. Glial cells (red) surround neurons (grey) and influence their electrical environment — a factor taken into account in a new model of neuronal behaviour.

A lion found in the Egyptian tomb of Maïa

Burial of a mummified lion at a dedicated site confirms this animal's once-sacred status.

Lions are mentioned by classical scholars and in pharaonic inscriptions as being among the sacred animals that were bred and buried in the Nile valley. And yet no specimens have been found in Egypt — until the excavation of the Bubasteion necropolis at Saqqara. Here we describe a complete skeleton, once a mummy, of a male lion (*Panthera leo*) that was discovered there, buried among the cats' catacombs¹ created during the last centuries BC and occupying the much older tomb of Maïa, wet-nurse to the king Tutankhamun (from the New Kingdom, fourteenth century BC). This important find at a site that was dedicated to the feline goddess Bastet (also known as Bubastis) confirms the status of the lion as a sacred animal during the Late and Greek periods.

Rock-cut tombs of the XVIIIth and XIXth dynasties have been excavated by the French Archaeological Mission of the Bubasteion (MAFB) on the site of Saqqara, cemetery of Memphis. These New Kingdom tombs have been partly reused as cats' catacombs and are located inside the Bubasteion precinct dedicated to the cat (once lioness) goddess Bastet during the last centuries BC (see supplementary information).

The tomb of Maïa (Fig. 1) was discovered by MAFB in 1996. It includes at least two levels: the chapel and the funerary apartments. The funerary apartments have no remains of the original burial and have been reoccupied by intrusive burials, first of humans and then of animals (mainly cats). The main funerary level is divided into three rooms connected by corridors. Room number one is roughly L-shaped and is supported by a rock-cut pillar, whereas room number two is smaller and roughly rectangular in shape, with a shaft leading to a lower and less important level. Room number three (4.5 m by 4 m) has been carefully hewn and prepared. The floors of these three rooms were covered with wooden coffins or sarcophagi and many decayed human and animal remains, with huge quantities of bones. The most numerous occupants were mummified cats, but stratigraphic features prove that the humans were deposited there beforehand, perhaps several centuries earlier.

While working in the main room of the funerary level in November 2001, we made the surprising discovery of a virtually complete, undisturbed skeleton



Figure 1 Sacred burial of a lion (*Panthera leo*). Left, King Tutankhamun and his wet-nurse, lady Maïa, on a relief carved in the chapel of her tomb. Right, skeleton of the once-mummified lion lying on the floor of room 3 in the funerary level of the tomb.

of a felid that was much larger than the cats found previously. It was a lion, lying on a rock (Fig. 1) with its head turned northwards and its body orientated to the east. The lion was surrounded by numerous animal bones and coffins that were partly destroyed and buried much later than the human occupant (see supplementary information).

The conservation of this *Panthera leo* specimen was excellent, except that the skull was partly crushed and the scapula and

femur had been damaged. Although there were no linen bandages to indicate that the body had been mummified, there is other evidence to indicate that it had — namely, the position of the skeleton, the presence of small and degraded fragments of tissue inside the cavity of the canine teeth, and deposits on and coloration of the bones that are similar to those of mummified cats discovered on the site.

At first sight, the general position of the



Figure 2 Skeletal features of the lion lying in Maïa's tomb. A scale drawing is shown of the skeleton of the *Panthera leo* specimen: blue, forelimbs (dark, right side; light, left side); yellow, hindlimbs (dark, right side; light, left side). Drawing courtesy of C. Callou/MAFB. © Hypogées. Inset, lower mandible of the *Panthera leo* found at Saqqara. Regrettably, it is impossible to give a precise age for the skeleton because of the poor dental state. The most remarkable dental abnormality is observed on the superior and lower canines. Their occlusal surfaces are perfectly smooth and very worn. Moreover, the dental crown is extremely low, ground down to just above the neck. Premolars and molars, upper and lower, are also markedly altered. It seems that this change results from extreme wear, because only the basal parts of roots of all the teeth remain. Only the chewing of very tough materials (such as pebbles) can explain such modifications of the teeth. Moreover, the aspect of the jaw is due to important and continuous pathologies, doubtless connected to deficiencies. This project was carried out with the permission of the Supreme Council of Egyptian Antiquities, Cairo. Photo, P. Chapuis/MAFB. © Hypogées.



skeleton is reminiscent of the position of the mummies of cats, even though the disposal of the hindlimbs is different. In the case of the lion, the forelimbs and the hindlimbs are stretched down along the front of the body, with the tail returned between the paws, but in the cats' mummies, the hindlimbs were tucked up against the pelvis with the tail curled up between the feet (Fig. 2). The bones are in strict connection, except the calvarium, which had fallen slightly behind, probably when the packaging disintegrated.

As the bone measurements are among the largest recorded for a male lion and outside the range of variation shown by females (C. Gross, personal communication), there is no doubt that this skeleton is from a male lion (see supplementary information). In contrast to results revealed by X-rays of many cat mummies^{1,2}, there is no indication that the animal was killed when young in order to be mummified and buried. It is likely that it died naturally.

Complete epiphyseal fusion of the skeleton and the very weak pulp chamber of the teeth show that they belong to an adult. The wear and pathology of the teeth (Fig. 2) suggest that the lion lived to be old and was kept in captivity. Seven ribs, on the same side of the chest, are fractured in the middle and have calluses formed on them; they may have broken as a result of a fall or a violent knock.

The existence of lions during the time of the pharaohs has been frequently described³, but to our knowledge this is the first complete skeleton to be found in Egypt (though some bones were found in Abydos⁴). The discovery should convey valuable information about the animal and its life (age, pathologies and accidents). Although buried in the tomb of a woman from the XVIIIth dynasty (about 1430 BC), the animal belongs to the later Bubasteion catacombs connected to the cult of animals that was so important in Late and Hellenistic Egypt (see supplementary information).

As a male, the mummified lion may have been considered as an incarnation of the god Mahes (Mysis)⁵, son of the goddess Sekhmet or Bastet. Although a single lion burial does not constitute a lion necropolis, this Memphite specimen confirms the existence of lions in Egypt in the last centuries BC. Lions are thought to have been bred in sanctuary precincts and buried in a sacred animal necropolis, including the Memphite necropolis (as suggested by demotic and Greek inscriptions⁶ and by some claws with evidence of mummification⁷).

Cécile Callou*§, Anaïck Samzun†§, Alain Zivie‡§

*MNHN et CNRS (UMR 5197), Paris, France

†INRAP, Paris, et UMR 7041, Nanterre, France

‡CNRS (UMR 8567), Paris, France

§Mission Archéologique Française du Bubasteion, Saqqara, Egypt

e-mail: az.hypogees@wanadoo.fr

1. Zivie, A. & Lichtenberg, R. in *Proc. VIIIth Int. Congress Egyptologists* Vol. 2, 605–611 (American University in Cairo Press, New York, 2003).
2. Armitage, P. & Clutton-Brock, J. *MASCA J.* 1, 185–188 (1980).
3. Houlihan, P. F. *The Animal World of the Pharaohs* (Thames & Hudson, London, 1996).
4. Boessneck, J. & von den Driesch, A. *Mitt. Deutschen Archäologischen Inst., Abteilung Kairo* 46, 86–87 (1990).
5. Zabkar, L. V. *Mysis (Miös, Mahes). Lexikon der Ägyptologie* (Wiesbaden) IV, 163–165 (1980).
6. Thompson, D. *Memphis under the Ptolemies* 29 n. 115 (Princeton Univ. Press, 1988).
7. Charron, A. *Bull. Soc. d'Égyptologie Genève* 21, 5–10 (1997).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Olfaction

Mosquito receptor for human-sweat odorant

Female *Anopheles* mosquitoes, the world's most important vector of *Plasmodium falciparum* malaria, locate their human hosts primarily through olfactory cues¹, but the molecular mechanisms that underlie this recognition are a mystery. Here we show that the *Anopheles gambiae* protein AgOr1, a female-specific member of a family of putative odorant receptors^{2,3}, responds to a component of human sweat. Compounds designed to activate or block receptors of this type could function as attractants for trapping mosquitoes or as insect repellents in helping to control *Anopheles* and other insect pests.

A member of the AgOr gene family, AgOr1, was expressed in an engineered neuron of the fruitfly *Drosophila*. The neuron failed to respond normally to odours because its endogenous odorant-receptor genes, *Or22a* and *Or22b*, had been deleted. AgOr1 was expressed in this mutant by using an *Or22a* promoter and the *GAL4-UAS* system⁴. We assayed the response of this neuron to individual odours by single-unit electrophysiology.

We found that AgOr1 confers a strong response to the odorant 4-methylphenol (Fig. 1a). As this compound is a component of human sweat that elicits an electrophysiological response from the antenna of female *A. gambiae*⁵, it may contribute to the anthropophilic host-seeking behaviour of this mosquito. This idea is supported by the fact that AgOr1 is expressed specifically in the olfactory tissue of only female mosquitoes, and its expression is downregulated after a blood meal² (the host-seeking behaviour of these mosquitoes is specific to the female and is reduced after blood-feeding⁶). Furthermore, 4-methylphenol increases the effectiveness of traps for the tsetse fly *Glossina morsitans morsitans*⁷.

We tested a second AgOr gene, AgOr2, and found a different odour-response spectrum (Fig. 1b). In contrast to AgOr1, AgOr2 confers a strong response to 2-methylphenol

but not to 4-methylphenol. There was no response in the case of the deletion mutant carrying no transgene (Fig. 1c).

These results indicate that this pair of AgOr genes encodes odorant receptors, and that the female-specific receptor AgOr1 may participate in the host-seeking behaviour of *A. gambiae*. As mosquito odorant receptors can function in *Drosophila* in the absence of other mosquito proteins, there could be a compatibility between odorant receptors and olfactory-receptor neurons from different species, which would allow the fruitfly to be used as an *in vivo* model for the study of odorant receptors derived from less genetically tractable insect species.

Anopheles mosquitoes transmit malaria and are responsible for the death of more than one million people each year. Our discovery that this mosquito possesses odorant receptors for particular components of human sweat means that different ligands could be screened for their activation or inhibition of these receptors, potentially

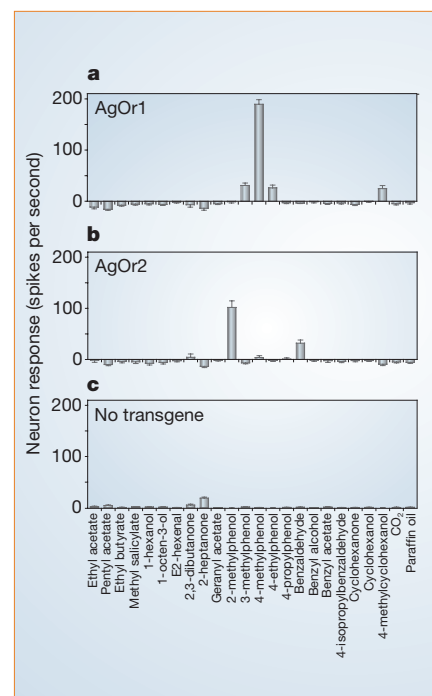


Figure 1 Identification of a mosquito odorant receptor that responds to a component of human sweat by expression in a *Drosophila* olfactory receptor neuron. **a–c**, Odour-response spectrum conferred by AgOr1 (**a**) and AgOr2 (**b**) on a *Drosophila* olfactory neuron carrying a deletion of its endogenous receptor genes (response of deletion mutant without transgenes is shown in **c**). Transgenic flies were of the genotype *w¹¹¹⁸; Δhalo/Δhalo*; *UAS-AgOr/Or22a promoter-Gal4* (refs 4, 8). Single-unit recordings were obtained⁴ from animals that were less than one week old. Liquid odour sources were diluted by 10^{−4} in paraffin oil; solid odour sources were diluted to 0.2 mg ml^{−1} in paraffin oil; CO₂ was administered as described in ref. 4. Responses were quantified by subtracting the number of spikes in 500 ms of spontaneous activity from the number in the 500 ms after the onset of odour stimulation (*n* = 12; error bars represent s.e.m.). AgOr cDNA clones were from *Anopheles gambiae sensu stricto* (G3 strain); embryos (provided by M. Benedict) were reared as described⁸.

leading to new, more effective insect traps and repellents.

Elissa A. Hallem*, A. Nicole Fox†‡, Laurence J. Zwiebel†, John R. Carlson*

*Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA
e-mail: john.carlson@yale.edu

†Department of Biological Sciences, Program in Developmental Biology and Center for Molecular Neuroscience, Vanderbilt University, Nashville, Tennessee 37235, USA

‡Present address: Division of Biology, California Institute of Technology, Pasadena, California 91125, USA

1. Takken, W. in *Offaction in Mosquito-Host Interactions* (eds Bock, G. R. & Cardew, G.) 302–312 (Wiley, New York, 1996).
2. Fox, A., Pitts, R., Robertson, H., Carlson, J. R. & Zwiebel, L. *Proc. Natl Acad. Sci. USA* **98**, 14693–14697 (2001).
3. Hill, C. A. *et al. Science* **298**, 176–178 (2002).
4. Dobritsa, A. A., van der Goes van Naters, W., Warr, C. G., Steinbrecht, R. A. & Carlson, J. R. *Neuron* **37**, 827–841 (2003).
5. Cork, A. & Park, K. C. *Med. Vet. Entomol.* **10**, 269–276 (1996).
6. Takken, W., van Loon, J. J. & Adam, W. J. *Insect Physiol.* **47**, 303–310 (2001).
7. Vale, G. A., Hall, D. R. & Gough, A. J. E. *Bull. Entomol. Res.* **78**, 293–300 (1988).
8. Gross, S. P., Guo, Y., Martinez, J. E. & Welte, M. A. *Curr. Biol.* **13**, 1660–1668 (2003).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Climatology

Rural land-use change and climate

Kalnay and Cai¹ claim that urbanization and land-use change have a major effect on the climate in the United States. They used surface temperatures obtained from NCEP/NCAR 50-year reanalyses (NNR) and their difference compared with observed station surface temperatures as the basis for their conclusions, on the grounds that the NNR did not include these anthropogenic effects. However, we note that the NNR also overlooked other factors, such as known changes in clouds and in surface moisture, which are more likely to explain Kalnay and Cai's findings. Although urban heat-island effects are real in cities, direct estimates of the effects of rural land-use change indicate a cooling rather than a warming influence that is due to a greater reflection of sunlight.

The NNR use upper-air observations to produce analyses of atmospheric fields every 6 hours by using four-dimensional data assimilation that capitalizes on available multivariate data. As a consequence of the procedures used, the NNR do not directly include local surface influences or observations. Therefore, land-use and urbanization effects may contribute to differences detected with high-quality, measured station surface temperatures.

In addition, the reanalyses do not include effects of the changing atmospheric compo-

sition on radiation, despite carbon dioxide concentrations in the atmosphere increasing from 318 p.p.m.v. in 1960 to more than 370 p.p.m.v. — an increase of about 17% — over this time. Neither do the reanalyses deal with changes in surface wetness. Increases in cloudiness and rainfall over the Mississippi River Basin have increased evaporation but decreased potential evapotranspiration². These trends have an important influence on the surface-heat balance.

The NNR did not include cloud information, and the depiction of clouds in the NNR is poor and the surface-heat budget has serious errors³. Detailed studies of the surface-heat budget and of why minimum temperatures are increasing at a faster rate than maximum temperatures reveal that the decreasing diurnal temperature range (DTR) is linked to a worldwide increase in cloud cover⁴. Clouds reduce DTR by sharply decreasing surface solar radiation and reducing radiative heat losses at night.

Processes involved in DTR, including radiation, surface fluxes of sensible and latent heat, and soil-moisture effects, have been investigated by using comprehensive measurements from the First ISLSCP (International Satellite Land Surface Climatology Project) Field Experiment (FIFE) in Kansas⁴. Changes in clouds, especially low clouds, largely determine the patterns of change of DTR. Soil moisture also decreases DTR by increasing cooling through daytime surface evaporation. Empirical relationships⁴ using 3-hourly weather observations extend the results globally to show that DTR varies inversely with cloud cover and precipitation on several timescales, particularly over the United States. The reported decreases¹ in DTR are therefore consistent with the observed increases in cloud cover.

Across the southern two-thirds of the eastern United States, the DTR peaks in spring and autumn, with minima in winter and mid- to late summer⁵. Changes in DTR are traceable to the lengthening growing season, especially on sunny days, indicating that the increases in vegetation and associated evapotranspiration are important.

However, a direct assessment of the effects of changes in land use and vegetation^{6,7} show that conversion of forests to crop land generally causes an increase in reflected sunlight that is greatest after the harvest in the autumn. The increased reflection results in a relative cooling, estimated to be in excess of 1 °C in autumn⁸, which is due to changes in land use rather than to warming¹. After the 1960s, the greatest land-use changes have been in the increase in crop land area in the midwestern United States and in reforestation in the northeast⁶. By contrast, urban heat-island effects are localized in cities, whose stations are not used in compilations of climate change. Also, changing snow cover contributes to the decrease

in DTR during winter in the United States⁸.

Changes in cloudiness and surface moisture are probably the main source of the discrepancies in trends found by Kalnay and Cai¹. The NNR omit these influences in computing the surface-heat budget, although they are critical for getting surface-air temperatures right. Influences by processes not in the NNR model (including urbanization and land-use change) will be included in variables whose observations are analysed, but not in those variables calculated from the model (including surface-air temperature).

Kevin E. Trenberth

National Center for Atmospheric Research,
PO Box 3000, Boulder, Colorado 80307, USA
e-mail: trenberth@ucar.edu

1. Kalnay, E. & Cai, M. *Nature* **423**, 528–531 (2003).
2. Milly, P. C. D. & Dunne, K. A. *Geophys. Res. Lett.* **28**, 1219–1222 (2001).
3. Trenberth, K. E., Caron, J. M. & Stepaniak, D. P. *Clim. Dyn.* **17**, 259–276 (2001).
4. Dai, A., Trenberth, K. E. & Karl, T. R. J. *Climate* **12**, 2451–2473 (1999).
5. Durre, I. & Wallace, J. M. J. *Climate* **14**, 354–360 (2001).
6. Bonan, G. B. J. *Climate* **14**, 2430–2442 (2001).
7. Myhre, G. & Myhre, A. J. *Climate* **16**, 1511–1524 (2003).
8. Durre, I. & Wallace, J. M. J. *Climate* **14**, 3263–3278 (2001).

Climate

Impact of land-use change on climate

Urbanization and other changes in land use have an impact on surface-air temperatures. Kalnay and Cai¹ report that the observed surface-temperature trend in part of the United States exceeds the trend in the NCEP/NCAR 50-year reanalysis (NNR) and conclude that changes in land use account for the difference (0.035 °C per decade according to their corrected values). Although land-use change may explain some of this discrepancy, the authors do not quantify the impact of the many changes in observational practice that occurred during the analysis period. Our findings indicate that these 'non-climatic' changes have a systematic effect that overwhelms the reported difference in trends and therefore calls Kalnay and Cai's central conclusion into question.

Historical archives kept at the National Climatic Data Center document many non-climatic changes in the area's observation stations. For example, from 1950 to 1999, over 25% of the stations switched from afternoon to morning observation schedules, imparting a gradual and systematic bias towards cooling to the area's temperature record². More than 75% of all stations experienced some change in instrumentation, and many were also relocated on one or more occasions.

The Baltimore Customs House, whose record is shown in Kalnay and Cai's Fig. 1,

is a station that has undergone several such changes (including instrumentation replacements in 1985, 1993 and 1998). Kalnay and Cai use raw data in their analysis but indicate that data adjustments for these non-climatic changes should result in a larger estimate of the impact of urbanization and other changes in land use.

To determine the effect of non-urban data adjustments, we performed an analysis identical to that of Kalnay and Cai, except that we used data from the US Historical Climatology Network (HCN) database³. The HCN is well suited for this purpose because it contains corrections that account for changes in observation time, instrumentation and location (adjustments for urbanization are also available, but were not used here because most HCN stations are in rural settings). We applied Kalnay and Cai's station-selection criteria to HCN, which yielded a set of 834 stations that are well distributed in their study area. We then used the authors' method to calculate the trend for the corrected HCN data for the period 1960–1999.

The resulting mean temperature trend in HCN (+0.224 °C per decade) exceeds Kalnay and Cai's observed temperature trend (+0.112 °C per decade). By the authors' reasoning, the difference between the HCN trend (+0.224 °C per decade) and the NNR trend (+0.077 °C per decade) is due to changes in land use. This trend difference (0.147 °C per decade) far exceeds their land-use value (0.035 °C per decade) and is ten times the size of the largest published urban estimate for the United States (0.015 °C per decade⁴). It also indicates that land-use change accounts for two-thirds of the warming over the past four decades. (The effect would have been even larger had a more urban network been used.)

In addition, this trend difference is decreasing over time. According to our calculations, the discrepancy between the corrected HCN trend and the NNR trend during the first two decades (0.202 °C per decade) is more than twice as large as during the past two decades (0.089 °C per decade).

These estimates seem improbable and indicate to us that the NNR trends are not accurate. We infer this in part because there is extensive evidence to support corrected HCN trends; they are spatially consistent with surface trends across international borders, with sea-surface temperature trends in adjacent oceans (which had no change in land use), and with tropospheric temperature trends derived from satellites and radiosondes⁵.

We are not aware of any evidence demonstrating the reliability of the NNR surface-temperature trends. Neither can we think of a reason why the land-use effect should have decreased by more than 50% during the study period. The decrease in the NNR land-

use estimate is particularly striking given the dramatic increase in temperature during the past two decades⁶ (+0.343 °C per decade in HCN).

Our results indicate that the NNR alone is not sufficient to identify a land-use impact, casting doubt on Kalnay and Cai's conclusions. However, their work does draw attention to an important issue that requires further investigation.

**Russell S. Vose, Thomas R. Karl,
David R. Easterling, Claude N. Williams,
Matthew J. Menne**

*National Climatic Data Center, Asheville, North
Carolina 28801, USA*

e-mail: russell.vose@noaa.gov

1. Kalnay, E. & Cai, M. *Nature* **423**, 528–531 (2003).
2. Karl, T. R., Williams, C. N., Young, P. J. & Wendland, W. M. *J. Clim. Appl. Meteorol.* **25**, 145–160 (1986).
3. Easterling, D. R. *et al.* United States Historical Climatology Network (US HCN) Monthly Temperature and Precipitation Data (Carbon Dioxide Inform. Anal. Cen., Oak Ridge Natl Lab., Tennessee, publication no. 4500, 1996).
4. Owen, T. W., Gallo, K. P., Elvridge, C. D. & Baugh, K. E. *Int. J. Remote Sensing* **19**, 3451–3456 (1998).
5. IPCC *Climate Change 2001: The Scientific Basis* (Cambridge Univ. Press, Cambridge, UK, 2001).
6. Hansen, J. *et al.* *J. Geophys. Res.* **106**, 23947–23964 (2001).

Cai and Kalnay reply — We do not deny the obvious importance of global warming and decrease in diurnal temperature range (DTR) due to greenhouse effects, which are present in both surface-station observations and the NCEP/NCAR 50-year reanalysis (NNR). Moreover, the NNR shows the largest warming trend over the past two decades, as reported in the surface-station data, suggesting that the NNR captures the dominant greenhouse-warming effect.

Our study¹ attributes the differences between the two data sets largely to land-use changes because the NNR is not subject to local surface influences. We deliberately used raw (unadjusted) surface observations and pointed out that the multiple non-climatic adjustments are uniformly positive, so our estimate should be considered as the lower bound of the effect of land-use changes. As we pointed out and Vose *et al.* confirm, adding these non-climatic adjustments to our lower-bound estimate does not alter the sign of the estimated land-use change effect but increases its magnitude.

Trenberth's comment that the reanalyses do not include the effects of the changing atmospheric composition seems to be based on the common misunderstanding that if the model used as a first guess does not have a carbon dioxide trend, for example, then the reanalysis may at best include only a 'watered-down' greenhouse-warming trend.

We showed by using an analytical study that the reanalysis can capture essentially the full strength of climate trends caused by the increase in greenhouse gases, even if this forcing is absent from the model used in the data assimilation (our unpublished data).

This is because the reanalysis assimilates atmospheric temperatures and other observations that are affected by greenhouse gases and other changes. We point out that, even though the model has no volcanic aerosols, a reanalysis can capture the atmospheric heating resulting from volcanic eruptions².

The fact that both station observations and the NNR exhibits a decrease in DTR reflects the impact of an increase in low-level clouds³. However, the surface observations show an even larger decrease in DTR, and we attribute the difference largely to land-use changes. This agrees with previous studies showing that urban effects also have a substantial impact on the decrease of DTR⁴.

The non-climatic adjustments can be added *a posteriori* to our estimate, leading to an upper-bound estimate of the impact of land-use changes. According to the calculations made by Vose *et al.*, the non-climatic adjustments to these raw station observations yield an averaged increase of 0.112 °C per decade. In other words, half of the averaged increase between 1960–1979 and 1980–1999 derived from the HCN data set (0.224 °C per decade) is the result of the non-climatic adjustments to the raw station observations.

Adding these non-climatic adjustments to our lower-bound estimate of the impact of land-use changes (0.035 °C per decade) yields 0.147 °C per decade. This upper-bound estimate is comparable to another study that also used the HCN data (0.12 °C per decade⁵). Therefore, the upper-bound estimate is not ten times the size of the largest published urban estimates for the United States.

We found that a decrease in the effect of total land-use change in 1960s–1970s to 1980s–1990s took place, primarily, over the rural stations. Reforestation, saturation of urban heat-island effects, and more regulated land-use changes could be leading factors resulting in such a decrease in land-use change. This decrease is independent of, and in no way contradicts, the "dramatic increase in temperature during the past two decades" because the NNR estimate also registers a larger increase in the daily mean surface temperature equal to 0.254 °C per decade over the past two decades, which is comparable with the estimate (0.343 °C per decade) derived from the HCN data set.

Ming Cai*, Eugenia Kalnay†

**Department of Meteorology, Florida State
University, Tallahassee, Florida 32306, USA
e-mail: cai@met.fsu.edu*

*†Department of Meteorology, University of
Maryland, College Park, Maryland 20742, USA*

1. Kalnay, E. & Cai, M. *Nature* **423**, 528–531 (2003).
2. Andersen, U., Kaas, E. & Alpert, P. *Geophys. Res. Lett.* **28**, 991–994 (2001).
3. Dai, A., Trenberth, K. E. & Karl, T. R. *J. Climate* **12**, 2451–2473 (1999).
4. Gallo, K. P., Easterling, D. R. & Peterson, T. C. *J. Climate* **9**, 2941–2944 (1996).
5. Kukla, G., Gavin, J. & Karl, T. R. *J. Climate Appl. Meteorol.* **25**, 1265–1270 (1986).

Substrate twinning activates the signal recognition particle and its receptor

Pascal F. Egea¹, Shu-ou Shan^{1,2}, Johanna Napetschnig¹, David F. Savage¹, Peter Walter^{1,2} & Robert M. Stroud¹

¹Department of Biochemistry and Biophysics, University of California at San Francisco, California 94143-2240, USA

²Howard Hughes Medical Institute, San Francisco, California 94143, USA

Signal sequences target proteins for secretion from cells or for integration into cell membranes. As nascent proteins emerge from the ribosome, signal sequences are recognized by the signal recognition particle (SRP), which subsequently associates with its receptor (SR). In this complex, the SRP and SR stimulate each other's GTPase activity, and GTP hydrolysis ensures unidirectional targeting of cargo through a translocation pore in the membrane. To define the mechanism of reciprocal activation, we determined the 1.9 Å structure of the complex formed between these two GTPases. The two partners form a quasi-two-fold symmetrical heterodimer. Biochemical analysis supports the importance of the extensive interaction surface. Complex formation aligns the two GTP molecules in a symmetrical, composite active site, and the 3'OH groups are essential for association, reciprocal activation and catalysis. This unique circle of twinned interactions is severed twice on hydrolysis, leading to complex dissociation after cargo delivery.

The SRP and its receptor (SR) target nascent proteins destined for secretion or membrane integration to the endoplasmic reticulum (ER) membrane in eukaryotes or to the plasma membrane in prokaryotes^{1,2}. The principal components of SRP and SR are conserved across all kingdoms. In prokaryotes the SRP is a ribonucleoproteinaceous complex consisting of a single 48 kDa GTPase, Ffh (the orthologue of the eukaryotic SRP54), and a 110-nucleotide 4.5S RNA (the orthologue of the eukaryotic 7SL RNA), whereas SR consists of a single GTPase called FtsY.

SRP54 associates with ribosomes near the peptide exit site³, and binds to signal sequences of membrane and secretory proteins as they emerge from the ribosome during translation⁴. SRP, with its ribosome and nascent polypeptide as cargo, then binds to SR (or FtsY)^{5,6}, which in turn is dynamically associated with the Sec61 (or SecYEG) translocon in the target membrane. Thus, SRP and SR act as molecular matchmakers to deliver the nascent polypeptide to the translocon. At the membrane, the ribosome and nascent polypeptide are released from SRP^{5,7,8} and bound instead by the translocon^{9–11}. SRP and SR reciprocally stimulate each other's GTPase activity, leading to dissociation of the complex^{12–14}.

Ffh comprises three domains, termed N, G and M¹⁵. The M domain contains the signal-sequence-binding site and provides a primary contact with 4.5S RNA, and is connected to the N and G domains by means of a flexible linker. The amino-terminal N domain and the GTPase G domain are structurally and functionally coupled in an arrangement that is universally conserved between all SRP GTPases¹⁶. The N domain is a four-helix bundle^{16,17} that changes its angle relative to the G domain depending on the nucleotide-bound state of the G domain¹⁸. The G domain is a Ras-like GTPase^{17,19} with an additional β - α - β - α insertion box domain (IBD) that is unique to the SRP subfamily of GTPases^{17,20} (Fig. 1a). The N and G domains of both Ffh and FtsY are sufficient for stable complex formation in the presence of the non-hydrolysable GTP analogues GMPPCP or GMPPNP. The structure of the complex reported here with GMPPCP elucidates a new mechanism by which complex formation activates both GTPases, and explains how subsequent GTP hydrolysis would break the interaction and cause dissociation of Ffh from FtsY. Each GTPase provides its own catalytic machinery *in cis* and is stimulated *in trans* by hydrogen bonds between the twinned GTPs.

The Ffh–FtsY heterodimer association

The X-ray structure of the catalytic core formed by the *Thermus aquaticus* SRP receptor (FtsY) and the N and G domains of the SRP protein (Ffh) in the presence of the non-hydrolysable GTP analogue GMPPCP was determined to 1.9 Å resolution (Table 1). The two GTPases form a quasi-two-fold symmetrical heterodimer through a continuous interface that surrounds the two bound GTPs and involves the G and N domains of both proteins (Fig. 1b, c). The primary interaction is between the G domains of FtsY and Ffh, and accounts for 2,500 Å² of the total 3,200 Å² buried surface area, whereas the N domains interact mostly with each other, and their

Table 1 Crystallographic data collection and refinement statistics

Data set	ALS 8.3.1 native
Wavelength (Å)	1.00
Resolution (last shell) (Å)	50–1.9 (1.97–1.9)
Total reflections	305,159
Unique reflections	47,344
Redundancy	6.3
Completeness (last shell) (%)	99.3 (97.9)
R_{sym} (last shell) (%)	7.7 (76.7)
I/σ (last shell)	16.9 (3.1)
Refinement statistics	FtsY–Ffh–GMPPCP complex
Reflections in working set (92.5% at 2σ)	41,806
Reflections in test set (7.5% at 2σ)	3,436
R_{cryst} (%)	20.6%
R_{free} (%)	23.9%
r.m.s.d. bonds (Å)	0.005
r.m.s.d. angles (°)	1.4
FtsY and Ffh	
Non-hydrogen protein atoms	2,091 and 2,183
Non-hydrogen ligand atoms	32 and 32
Ions (magnesium)	1 and 1
Solvent molecules	387
Average B-factors (Å ²)	
B-Wilson from data	13
Overall protein	22 and 22
Overall G domains	16 and 18
Overall N domains	32 and 35
Ligand atoms (including Mg ²⁺ ion)	14 and 14
Solvent molecules	31

r.m.s.d. is the root mean square deviation from ideal geometry. $R_{\text{sym}} = \sum_i \sum_j |I_{hkl,i} - I_{hkl,j}| / \sum_i \sum_j I_{hkl,i}$, where $I_{hkl,i}$ is the average intensity of the multiple hkl,i observations for symmetry related reflections. $R_{\text{cryst}} = \sum_i |F_o - F_c| / \sum_i F_o$. F_o and F_c are observed and calculated structure factors, R_{free} is calculated from a set of randomly chosen 7.5% reflections (2σ), and R_{cryst} is calculated over the remaining 92.5% of reflections (2σ).

interaction surface accounts for 700 Å² (Fig. 2), suggesting that complex formation is driven primarily by the pairing of G domains.

The interface is stabilized by extensive pairing interactions between Ffh and FtsY (Fig. 2a): 39 residues of FtsY and 34 residues of Ffh form 21 hydrogen bonds (10 direct and 11 water-mediated) and 139 van der Waals contacts. Three major areas of the interface are contributed to by residues interacting with each other across the quasi-two-fold axis (Fig. 2b, orange arrows, and Fig. 2c): (1) residues from motif II (Fig. 1) form extensive van der Waals contacts; (2) the N-terminal ends of the α3 helices come into close contact at the position of two conserved glycine residues; and (3) the conserved 'ALLEADV' motifs in the N domains that pack tightly against the base of the G domain are linked by means of an extensive network of interactions with α3 and α4 helices of the binding partner. There are no cavities in the interface with the single exception of the region localized between the two CH₂ groups of the GMPPCP molecules (see below). The sequestration of bound GTPs from solvent rationalizes the observation that there is no dissociation of GTP from the complex without GTP hydrolysis and concomitant dissociation of the complex^{21,22}.

An independent biochemical analysis of the Ffh–FtsY interaction agrees with the structure of the complex. This analysis was performed with *Escherichia coli* FtsY, which is highly homologous to *T. aquaticus* FtsY and on which most functional studies have been performed. Mutagenesis was carried out on conserved surface residues. Each of 40 mutant proteins was purified and assayed *in vitro*. A large proportion of the mutant FtsY proteins severely compromise the reciprocally stimulated GTPase reaction of FtsY with Ffh. The second-order rate constant (k_{tot}) for the reaction GTP·Ffh + FtsY·GTP → products, which includes both complex

formation and activation of GTP hydrolysis, is reduced 5–10³-fold in 25 out of the 40 mutants (Fig. 3a, black bars). The deleterious effects are not due to improper folding or impaired GTP binding of the mutant proteins, as the basal GTPase activities of most (36 of 40) of the mutant FtsY proteins are not significantly decreased (less than twofold) compared with wild-type FtsY (see Supplementary Table S1). The remaining four mutations decrease the basal GTPase activity to a greater, albeit still modest degree (3–8-fold; Supplementary Table S1). Nucleotide binding was not significantly affected in any of the mutants (data not shown).

When mapped onto the homologous residues in the structure of *T. aquaticus* FtsY in complex with Ffh (Fig. 3b), all of the mutants that have deleterious effects (red), with the exception of FtsY(K399(204)A) (where the number in parentheses relates to the homologous *T. aquaticus* sequence), lie within the FtsY–Ffh interface. This includes residues within conserved motifs in the GTPase site, the 'closing loop' that packs against the guanine base, and the N–G domain interface. All of the neutral mutants (Fig. 3b, yellow) lie on the opposite side, pointing into solvent and away from the interface. Thus, the extensive interaction surface seen in the structure is functionally important for the stability and/or activity of the complex.

Conformational change that activates both GTPases

Superposition of the structures of FtsY·GMPPNP and Ffh·NG·GMPPNP on the heterodimeric complex shows that, on complex formation, the highly conserved motifs II (133–145; IBD loop) and III (187–192) in each G domain undergo major conformational rearrangements as they seal the upper part and lateral entrance of the catalytic sites from solvent water (Fig. 4a, b). As a consequence, the two IBD loops each bring three key side chains into the active

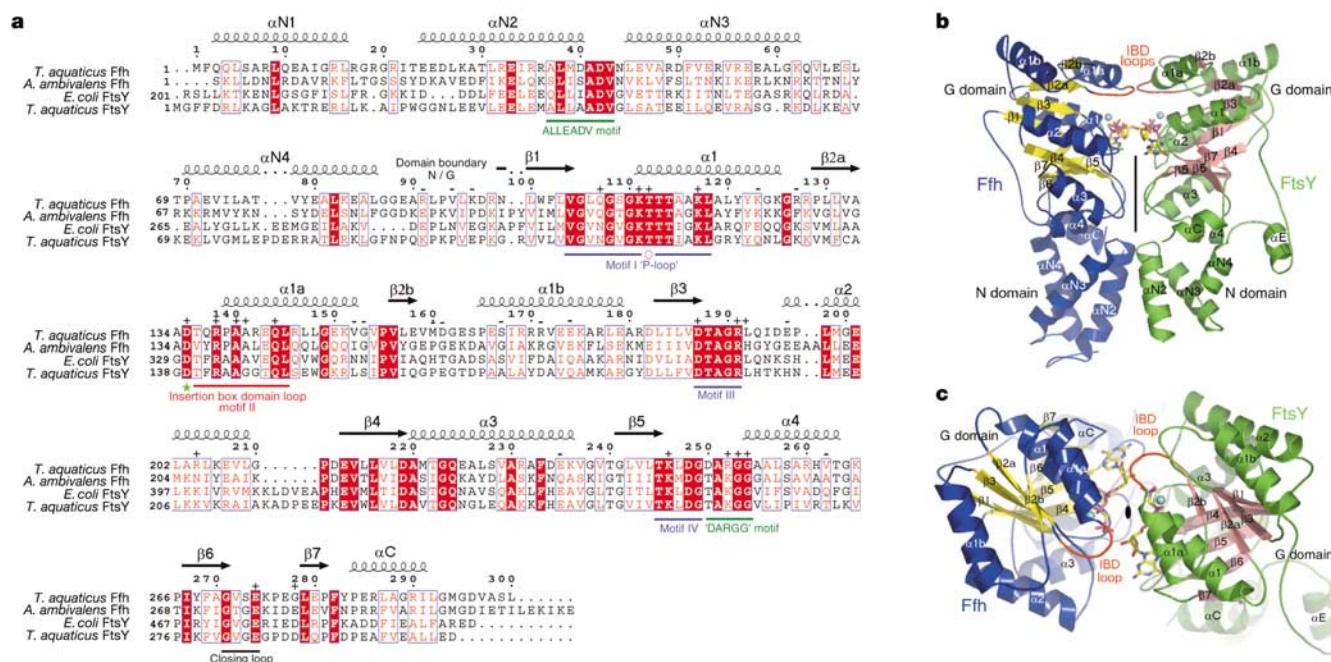


Figure 1 Sequence alignment of SRP GTPases of known structure, and structure of the FtsY–Ffh heterodimer. **a**, Sequences of Ffh (*T. aquaticus* and *Acidianus ambivalens*) and FtsY (*E. coli* and *T. aquaticus*). Conserved regions (red) include motifs I (P-loop), II, III and IV, the closing loop, the ALLEADV and DARGG motifs, T112 (red circle) and D135 (asterisk). Residues of similar chemical nature (boxed in blue) and secondary structure elements are indicated. Mutations (see Fig. 3a) affecting GTPase activity (+) and neutral mutations (–) are also indicated. Numbering in the text is the residue and number within its own sequence, followed in brackets by the residue if different and the number of the

homologous residue in *T. aquaticus* Ffh, which we use as a reference for all SRP GTPases. *E. coli* FtsY contains an additional N-terminal domain, of unknown function, that is absent in *T. aquaticus* FtsY. **b**, **c**, The heterodimer shown in two orientations. Blue and yellow colouring of Ffh indicate α-helices and β-strands; the corresponding structures in FtsY are shown in green and pink. The quasi-two-fold axis is indicated in black. The two bound GMPPCP molecules (yellow sticks) coordinated to two Mg²⁺ ions (cyan spheres), and the IBD loops (red), are shown. See Supplementary Information for a movie of the structure.

site of their respective protein (see below) and also contribute to the heterodimerization interface.

A highly conserved interface consisting of the ALLEADV motif within the N domain and the 'DARGG' motif within the G domain acts as a fulcrum that allows readjustment of the relative position of the N and G domains. Both N domains also change conformation on complex formation, bending towards the quasi-two-fold axis and approaching closest contact at the level of the N–G domain interface (Fig. 2b, grey arrows). These changes are coupled to the G domains through extensive interactions across the N–G interface by means of the DARGG motifs. As a consequence, conserved motif IV is brought into closer contact with the guanine base. In particular the specificity determinant D258(248) in FtsY optimizes hydrogen bonding with the N2 and N3 amino groups (the average hydrogen bond distance is 2.7 versus 3.4 Å in the complex of FtsY–Ffh versus free FtsY). This rearrangement explains the observation that FtsY acquires substantial nucleotide specificity only on its association with Ffh²³.

Substrate twinning at a composite active site

Trans-substrate interactions

The two GTPase sites are paired to form a composite active site at the interface between FtsY and Ffh. The extended analogues lie

approximately perpendicular to the quasi-two-fold axis. Thus, the two GMPPCPs are twinned to face each other in a head-to-tail manner (Figs 1b, c and 2b, red arrows). To our knowledge, such a direct nucleotide–nucleotide association is unprecedented. The two nucleotides are sequestered from bulk solvent and bound such that the γ -phosphate of each GMPPCP accepts a 2.6 Å hydrogen bond from the 3'OH of the ribose ring of the other GMPPCP (Fig. 5). The ribose sugar pucker is C3'-endo in both cases, providing ideal geometry of the OH donor, and makes the active site reciprocal in atomic detail. This is the primary contact between active sites across the heterodimer interface.

To test the role of the *trans*-substrate interactions between the ribose 3'OH groups and the γ -phosphate oxygens, we determined the effect of removing the hydrogen bond donor by replacing the 3'OH of each GTP with 3'H. This was facilitated by the use of *E. coli* mutant Ffh and FtsY, which have altered nucleotide specificity for XTP (xanthosine 5'-triphosphate) instead of GTP (Ffh(D251(248)N) and FtsY(D449(248)N))^{14,23}. Use of the mutants allows selective occupation of one of the GTPase sites with GTP or GTP analogues, and permits comparisons of their ability to stimulate the rate of XTP hydrolysis on the partner.

Figure 3c shows the stimulation of XTP hydrolysis on mutant FtsY(D449(248)N) by Ffh molecules bound with GTP analogues.

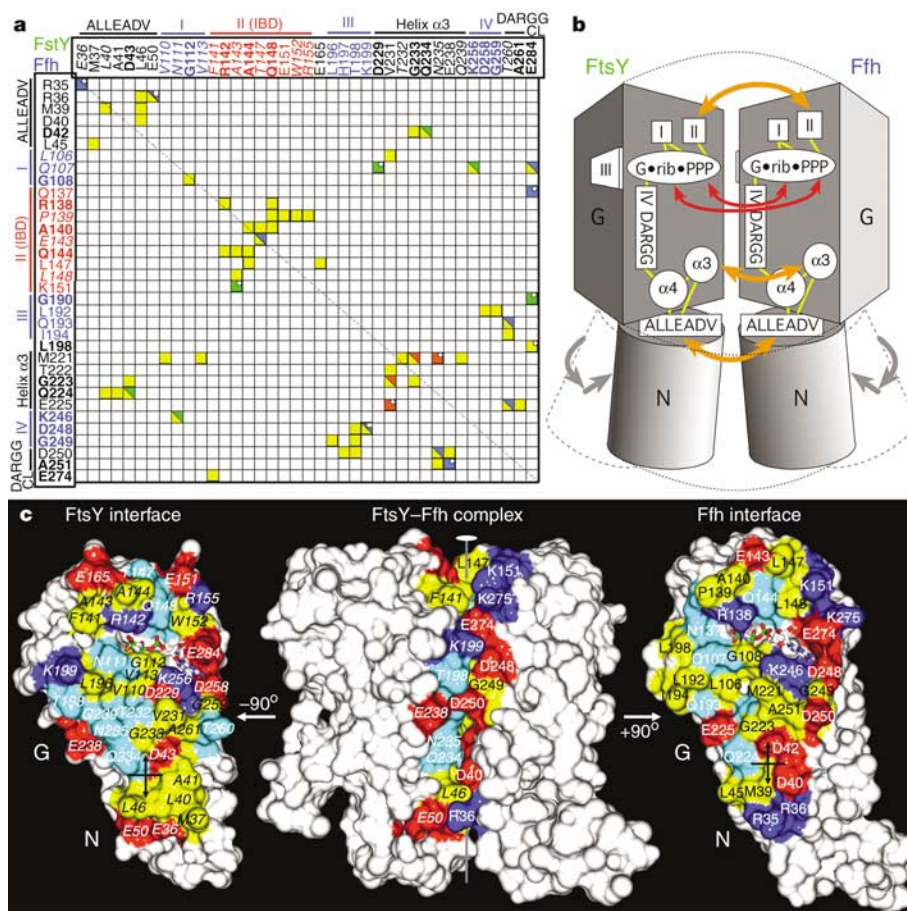


Figure 2 The heterodimerization interface. **a**, Inter-residue contacts represented in a matrix. Van der Waals contacts (yellow) and hydrogen bonds (blue for side chain to side chain, green for side chain to backbone, and red for backbone to backbone) are shown. Water-mediated hydrogen bonds or ionic interactions are represented as white dots and white squares, respectively. Residue numbers and the conserved motifs are indicated. Residues indicated in bold are strictly conserved in both proteins. Symmetry across the diagonal emphasizes the quasi-two-fold symmetry. **b**, Interactions in the interface (orange

arrows), nucleotide twinning (red arrows) and rearrangements (grey arrows). Motifs I, II, III and IV are involved in nucleotide-binding and/or catalysis. The helices α 4 act as adaptors between the N and G domains. **c**, Surface complementarity. The complex (middle) is opened -90° (FtsY, left) and $+90^\circ$ (Ffh, right). GMPPCP (in sticks) and side chains (positive, blue; negative, red; hydrophobic, yellow; hydrophilic, cyan) are mapped. The cross shows a point of contact on the quasi-two-fold axis (vertical). FtsY residues are italicized whereas Ffh residues are not.

We followed the reaction $\text{XTP} + \text{FtsY(D449(248)N)} + \text{Ffh-NTP} \rightarrow \text{XDP} + \text{P}_i$ (where NTP denotes GTP or GTP analogues and P_i denotes the phosphate anion), which includes effects on both Ffh–FtsY complex formation and activation of XTP hydrolysis. When Ffh is bound with 3'dGTP (Fig. 3c, filled circles) rather than GTP (triangles), the rate constant of this reaction is reduced 420-fold. In contrast, replacing the 2'OH of GTP with 2'H has only a twofold effect (Fig. 3c, open circles).

The same result holds for the reciprocal reaction; that is, the stimulation of XTP hydrolysis on mutant Ffh(D251(248)N) by FtsY molecules bound with GTP analogues ($\text{XTP-Ffh(D251(248)N)} + \text{FtsY-NTP} \rightarrow \text{XDP} + \text{P}_i$) (Fig. 3d). Replacing the 3'OH of GTP by 3'H reduced the rate of the reaction by 140-fold (Fig. 3d, closed circles versus triangles), whereas replacement of the 2'OH by 2'H had less than a twofold effect (open circles). Thus the 3'OH of the GTP bound in each active site has a crucial role, contributing between 3.0 and 3.7 kcal mol⁻¹ to binding and reciprocal GTPase activation.

A water molecule between the γ -phosphate of GMPPCP in FtsY and the α -phosphate of GMPPCP in Ffh also contributes to the interaction between nucleotides. The two CH₂ groups (7.1 Å between carbon atoms) and the two hydrophobic exo-faces of the ribose rings form a hydrophobic cavity with no electron density in

it. If GTP were bound, a water molecule(s), an ion, or a charged or polar side chain could fill the cavity, bridging oxygens of the β - and γ -phosphates. This additional interaction might contribute further to catalysis.

Interactions between active site and substrate

Each active site provides catalytic groups for the bound nucleotide *in cis*. In each GMPPCP one oxygen of the β -phosphate and one oxygen of the γ -phosphate is coordinated by a single Mg²⁺ ion in octahedral coordination with three water molecules and the OH group of a strictly conserved threonine residue from motif I (the 'P'-loop; T112 in Ffh and T116(112) in FtsY) (Fig. 5b). One water molecule in each GTPase lines up opposite the respective β – γ -phosphate bonds in ideal attacking position for hydrolysis (Fig. 5a). These were not present in the separate FtsY-GMPPNP or Ffh-GMPPNP structures²⁴. Each donates a hydrogen bond (2.6 Å) to a strictly conserved aspartate (D135 in Ffh and D139(135) in FtsY) from motif II (IBD loop; Fig. 5). These two aspartate residues also accept a hydrogen bond from one water molecule that coordinates the Mg²⁺. These aspartates are specific to the SRP family of GTPases and are the key, previously unrecognized catalytic groups that are only brought into correct orientation in the complex.

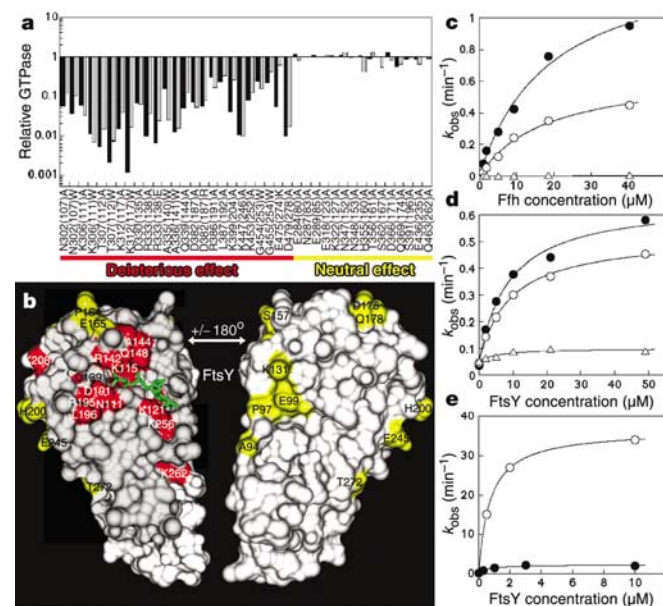


Figure 3 Functional analysis of conserved residues and the 3'OH of GTP. **a**, The effect of FtsY mutations on the stimulated GTPase reaction. The dark and grey bars are the k_{tot} and k_{cat} values, respectively, for the mutant relative to wild-type FtsY. Red denotes mutants with deleterious effects whereas yellow denotes neutral mutants. **b**, Mutants in *E. coli* FtsY (coloured as in **a**) mapped onto the corresponding surface of complexed *T. aquaticus* FtsY. The interface (grey, left) with the bound nucleotide (green sticks) and the solvent-accessible side (white, right) are shown. **c, d**, Replacing the 3'OH of each GTP with 3'H compromises Ffh–FtsY association and reciprocal stimulation. XTP hydrolysis from XTP-specific mutants FtsY(D449(248)N) (**c**) or Ffh(D251(248)N) (**d**) were determined with the other GTPase active site bound by GTP (filled circles), 2'dGTP (open circles) or 3'dGTP (triangles). First-order fits to the linear portion of the concentration dependences gave apparent second-order rate constants (k_{tot}) for XTP hydrolysis from FtsY(D449(248)N) (**c**) of 6.0×10^4 , 2.6×10^4 and $1.4 \times 10^2 \text{ M}^{-1} \text{ min}^{-1}$ with Ffh bound by GTP, 2'dGTP and 3'dGTP, respectively, and second-order rate constants for XTP hydrolysis from Ffh(D251(248)N) (**d**) of 7.2×10^4 , 5.4×10^4 and $5.3 \times 10^2 \text{ M}^{-1} \text{ min}^{-1}$ with FtsY bound by GTP, 2'dGTP and 3'dGTP, respectively. **e**, The stimulated GTPase reaction of mutant FtsY(D330(135)A) (filled circle) compared with the reaction of wild-type FtsY (open circle).

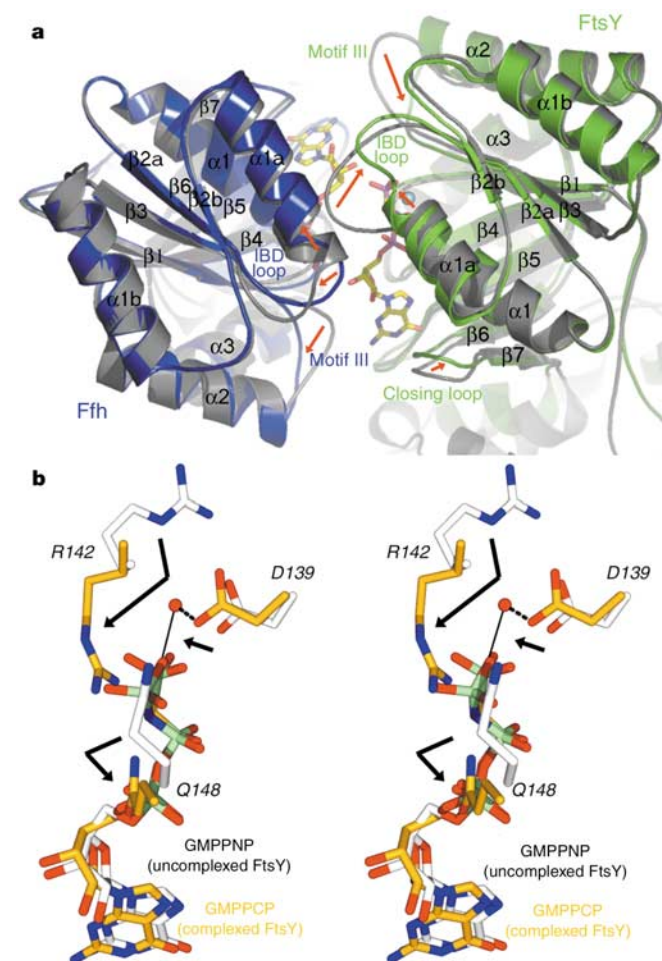


Figure 4 Conformational rearrangements on formation of the FtsY–Ffh complex. **a**, Changes in the IBD and motif III regions. The structure of the GMPPCP-bound FtsY–Ffh complex (FtsY, green; Ffh, blue) is superposed on the GMPPNP-bound structures of FtsY and Ffh (grey). Red arrows highlight the rearrangements in particular in the helices $\alpha 1a$ from the IBs. **b**, Structures of FtsY–GMPPCP from the FtsY–Ffh complex (orange) and free FtsY–GMPPNP (white) are superposed and show the changes (arrows) in the relative positions of the conserved residues D139(135), R142(138) and Q148(144).

Complex formation also brings R138 and Q144 of the Ffh IBD loop, and its conjugate pair of R142(138) and Q148(144) of the FtsY IBD loop, from far away into catalytic position in close contact with the α -, β - and γ -phosphates of the GMPPCPs (Fig. 5b). These residues act to position the nucleotide for hydrolysis, and contribute to the electrostatic balance within the catalytic site where the negatively charged phosphate groups of the twinned GMPPCPs converge. Upon hydrolysis an extra negative charge is developed in each site, compromising this balance and destabilizing the complex. This notion is supported by the mechanism proposed for ABC transporters²⁵, in which cooperative hydrolysis of two ATPs in a symmetrical homodimer drives domain dissociation, which in turn drives transport. However, in these structures the ATP substrates are far from each other and the mechanisms of coupling hydrolyses are different.

Additional symmetrical interactions involve close van der Waals contacts of side chains of R138 of Ffh with R142(138) and Q148(144) of FtsY and, reciprocally, of R142(138) of FtsY with R138 and Q144 of Ffh (Fig. 5b). In contrast, other conserved residues interact in ways that are not quite symmetrical. A hydrogen bond forms between the side chain of N111(Q107) of FtsY and the ribose 3' O of the GMPPCP bound to Ffh. This is matched—but not mirrored—by a van der Waals contact between Q107 of Ffh and the ribose ring of the GMPPCP bound to FtsY. Asparagine is strictly conserved in this position in all FtsY sequences and their eukaryotic orthologues; correspondingly, glutamine is conserved in all Ffh sequences and their eukaryotic orthologues. These two side chains are the only ones that interact with the opposing substrate. By breaking symmetry, they may be important in specifying the order of the two hydrolysis reactions in the SRP–SR heterodimer.

Consistent with the crucial role of the motif II loop in activation of the GTPases, mutations of residues in motif II (D135, R138, A140, A141 and Q144) all have deleterious effects (Fig. 3a). Figure 3e shows the effect of the D330(135)A mutation in *E. coli* FtsY. The first-order rate constant for the reaction $\text{GTP} \cdot \text{FtsY} \cdot \text{Ffh} \cdot \text{GTP} \rightarrow \text{products}$ (k_{cat} ; Fig. 3a, grey bars), which measures the activation of GTP hydrolysis reaction after the complex is formed, is reduced 18-fold for FtsY(D330(145)A) compared with wild-type FtsY (Fig. 3e, compare filled with open circles). As the two active sites hydrolyse bound GTPs with the same rate constant in the complex¹⁴,

inactivation of only one of the GTPases would reduce the observed rate constant by only twofold. Thus, the large effect of the D330(135)A and other mutations (Fig. 3a, red) indicates that the activation of both GTPase sites is compromised, even when mutations are confined to one of the two GTPases. The proper assembly of the composite active site is therefore highly cooperative *in trans* between the two proteins.

Discussion

The structure of the catalytic core of the SRP and SR complex suggests a unique mechanism by which nucleotide binding and hydrolysis drive complex formation and disassembly. The two proteins associate by means of an extensive interaction surface, which spans both the N and G domains, and undergo major conformational changes relative to the free proteins. Of the 40 mutations of surface residues that we made, 25 map to the interface and each shows a pronounced deleterious effect on complex formation. This indicates that the conformation of the protein in the complex is stable only when all required interactions are made. Thus, complex formation involves a highly cooperative and extensive network of interactions.

The closest related structures to the SRP GTPases include the homodimeric ATPase nitrogenase²⁶. However, a model of the FtsY–Ffh complex²⁰ based on this failed to predict the N domain or mechanistic interactions. In contrast to most other GTPases, major rearrangements occur in Ffh and FtsY only as they associate in their GTP-bound forms. As free proteins, there are only subtle differences between the GTP- and GDP-bound states²⁴. By contrast, other GTPases undergo marked rearrangements in response to GDP and GTP binding (for example, movements of 'switch regions' in motif II and motif III of Ras) and little further rearrangement takes place on binding their associated GTPase-activator proteins (GAPs).

The FtsY–Ffh complex forms a composite active site at the interface. In this site, many catalytic residues interact both with the substrate *in cis* and with residues across the interface *in trans*. As a result, complex formation and GTPase activation are highly coupled. Consistently, mutation of catalytic residues also impairs complex formation. Because of the composite nature of the active site, activation of each GTPase site is also highly coupled to

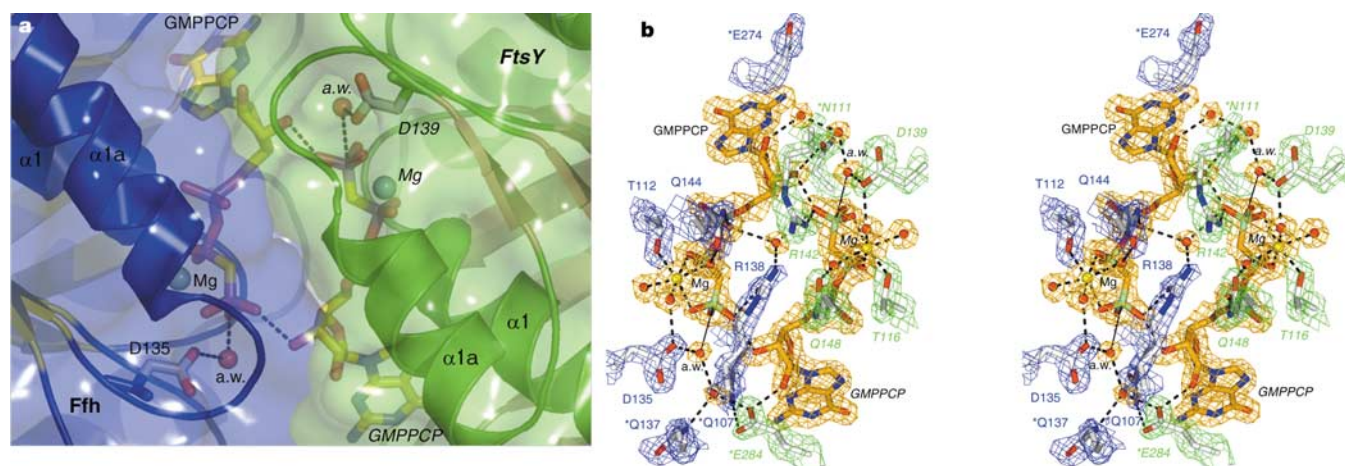


Figure 5 The composite catalytic site with the twinned substrates and essential residues. **a**, The twinned substrates showing symmetrical hydrogen bonds between the 3' OH ribose of one GMPPCP and the γ -phosphate of the other GMPPCP. The two attacking waters (a.w.) are aligned for nucleophilic attack. Protein surfaces are blue (Ffh) or green (FtsY) transparent envelopes. **b**, Stereo view of the twinned catalytic site. The $2F_o - F_c$ electron density map (2.1σ) corresponding to Ffh (blue), FtsY (green) and ligands

(orange). Residues in Ffh (blue characters) and FtsY (green italic characters) are shown; asterisks indicate residues breaking the pseudo-symmetry. Black dashed lines indicate hydrogen bonds. Orientations in **a** and **b** are identical. See Supplementary Information for a movie showing the catalytic residues, the attacking waters and the twinned nucleotides.

activation of the other, as revealed by single mutations in the active site of one GTPase that inactivate both GTPases.

The structure suggests a unique activation mechanism for the SRP family of GTPases. Conformational rearrangements on complex formation bring catalytic residues in the IBD loop into the active site and align them with respect to the bound substrate. In particular D135 activates the attacking nucleophilic water molecule, and R138 and Q144 interact with the β - and γ -phosphate groups, and thus may contribute to stabilization of the transition state. The only catalytically important interactions between the GTPases are formed by the twinned substrates themselves, which further stabilizes the negative charge on the γ -phosphate groups. In contrast, the active sites of most GTPases become complemented on interaction with their respective GAPs through insertion of a missing catalytic residue, such as the 'arginine finger'. The only other known exceptions are the Ran²⁷ and G α GTPases^{28–30}, which on activation align key catalytic residues *in cis*.

The extensive interactions with the γ -phosphate group, both from active-site residues *in cis* and from the twinned substrate *in trans*, explain why complex formation is GTP-dependent and why GTP hydrolysis leads to complex dissociation. Hydrolysis of GTP releases the γ -phosphate, and therefore severs the connections with both the substrate and active-site residues. Electrostatic repulsion between the released phosphate and GDP might assist further in dissociating the partners. The most basic requirement for SRP-dependent unidirectional targeting, namely the obligatory coupling of the formation of a tight SRP–SR complex with its subsequent disassembly, is therefore elegantly explained by these structural and functional analyses.

Note added in proof: A structure of the complex reported here in a different crystal form has been determined independently³¹. No information was shared prior to publication. □

Methods

Protein preparation

Full-length FtsY (residues Met 1 to Asp 304) and the N and G domains of Ffh (residues Met 1 to Leu 300) from *T. aquaticus* were expressed in BL21(DE3)–Rosetta *E. coli* cells, and subcloned in the pET28b vector as N-terminal hexa-histidine fusions. Proteins were expressed at 37 °C in LB medium. Soluble cellular extracts were heated at 70 °C for 30 min and clarified by centrifugation. FtsY and Ffh-NG were purified by metal affinity chromatography followed by gel filtration. The thrombin-cleavable histidine tags were removed. The FtsY–Ffh-NG binary complex was formed by incubation of stoichiometric amounts of FtsY and Ffh-NG in the presence of a twofold excess of the non-hydrolysable GTP analogue GMPPCP. The complex was purified by gel filtration and concentrated for crystallization. Expression plasmids for mutant FtsYs were constructed from wild-type *E. coli* FtsY(47–497). Truncation of the first 46 amino acids improves expression and has no discernible effect on the interaction with Ffh or on the GTPase activity of FtsY alone or complexed to Ffh. Mutant FtsYs were expressed and purified as for wild type²².

Crystallization and structure determination

The complex was crystallized by vapour diffusion at 4 °C using the Nextal crystallization pre-filled screening suites and tools (Nextal Biotechnologies). Crystals grew in 10% PEG 8,000 and 100 mM HEPES, pH 7.5, or 10% PEG 20,000 and 100 mM MES, pH 6.5, and were optimized at 4 °C and at 20 °C combining seeding and additive screening. Crystals were flash-frozen using mother liquor supplemented with 25% glycerol. Diffraction data collected at beamline 8.3.1 at the Advanced Light Source (ALS) at LBNL were integrated and scaled with DENZO and SCALEPACK³². The space group is $P2_12_1$ with unit cell dimensions $a = 75.0$ Å, $b = 83.7$ Å, $c = 94.0$ Å with one heterodimer per asymmetrical unit. The structure was solved by molecular replacement using the crystal structures of Ffh-NG and FtsY bound to GMPPNP as search models in AmoRe³³. Refinement in CNS³⁴ and model building in Moloc³⁵ yielded the structure consisting of FtsY (residues Asn 27 to Glu 303), Ffh (residues Gly 19 to Gly 295), two GMPPCP molecules and 387 water molecules. All residues are in the most allowed regions of the Ramachandran diagram. On complex formation only, there is enhanced hydrolysis of the first 25 residues of FtsY as previously noted³⁶ and as also seen here by mass spectrometry. These residues are removed perhaps by self-cleavage within the sequence G24–G25–N26. Before complex formation this variable region of the N domain of FtsY–GMPPNP is relatively unstructured. Whether this is important functionally is not yet known. The first 18 residues of the NG domain of Ffh are present but poorly ordered in the structure. Figures were prepared with PyMol³⁷ and DINO³⁸.

Kinetic analysis of the Ffh–FtsY interaction

Stimulated GTP hydrolyses by *E. coli* Ffh and FtsY were followed by monitoring

radioactive P_i release²³. An approximately twofold excess of 4.5S RNA over Ffh was present to facilitate the Ffh–FtsY complex formation. The reciprocally stimulated GTPase reactions were determined in multiple turnover experiments ($[GTP] > [Ffh + FtsY]$). Varying amounts of wild-type or mutant FtsY were added to a small, fixed amount of Ffh (0.05–0.1 μ M) in the presence of a saturating amount of GTP (100 μ M) to occupy both GTPase active sites ($K_d^{GTP} = 0.39$ for Ffh; $K_d^{GTP} = 14$ μ M for FtsY). Nucleotide affinity of FtsY was not significantly affected by any of the mutations reported here. The concentration-dependence of FtsY was analysed²² to obtain rate constants k_{tot} and k_{cat} for the summed GTP hydrolyses. k_{tot} is identical to the operationally defined second-order rate constant referred to as k_{cat}/K_m in previous work²².

XTP hydrolysis from mutant FtsY(D449(248)N) was determined in single-turnover reactions ($[XTP] < [E]$) with a fixed, subsaturating amount of FtsY(D449(248)N) (0.2 μ M) with respect to XTP ($K_d^{XTP} = 260$ μ M)²⁴, and varying amounts of Ffh. A 50 μ M concentration of GTP or GTP analogues was present to selectively saturate the GTPase site in Ffh ($K_d^{GTP} = 0.39$ μ M; $K_d^{3'dGTP} = 0.63$ μ M; $K_d^{2'dGTP} = 0.86$ μ M). The reaction $XTP + FtsY(D449(248)N) + Ffh-NTP \rightarrow XDP + P_i$ was followed by radioactive P_i release from XTP. The apparent second-order rate constant for the reaction (K_{tot}^{app}) was obtained from the slope of the initial linear portion of the dependence on concentration of Ffh.

The reciprocal reaction, XTP hydrolysis from mutant Ffh(D251(248)N), was determined in single-turnover experiments with a fixed, saturating amount of Ffh(D251(248)N) (2 μ M) with respect to XTP ($K_d^{XTP} = 0.31$ μ M; S.-o.S. and P.W., unpublished observations) and varying amounts of FtsY. A concentration of 200 μ M GTP or GTP analogues (NTPs) was present to selectively occupy the GTPase site in FtsY ($K_d^{3'dGTP} = 52$ μ M; $K_d^{2'dGTP} = 32$ μ M). Thus, the reaction $XTP-Ffh(D251(248)N) + FtsY-NTP \rightarrow XDP + P_i$ was followed. The rate constant for this reaction, k_{tot} , was obtained from the slope of the initial linear portion of the dependence on the concentration of FtsY.

Received 25 August; accepted 25 November 2003; doi:10.1038/nature02250.

- Keenan, R. J., Freymann, D. M., Stroud, R. M. & Walter, P. The signal recognition particle. *Annu. Rev. Biochem.* **70**, 755–775 (2001).
- Stroud, R. M. & Walter, P. Signal sequence recognition and protein targeting. *Curr. Opin. Struct. Biol.* **9**, 754–759 (1999).
- Pool, M. R., Stumm, J., Fulga, T. A., Sinning, I. & Dobberstein, B. Distinct modes of signal recognition particle interaction with the ribosome. *Science* **297**, 1345–1348 (2002).
- Rinke-Appel, J. *et al.* Crosslinking of 4.5S RNA to the *Escherichia coli* ribosome in the presence or absence of the protein Ffh. *RNA* **8**, 612–625 (2002).
- Gilmore, R., Blobel, G. & Walter, P. Protein translocation across the endoplasmic reticulum. I. Detection in the microsomal membrane of a receptor for the signal recognition particle. *J. Cell Biol.* **95**, 463–469 (1982).
- Meyer, D. I., Krause, E. & Dobberstein, B. Secretory protein translocation across membranes—the role of the 'docking protein'. *Nature* **297**, 647–650 (1982).
- Walter, P. & Blobel, G. Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum. *Proc. Natl Acad. Sci. USA* **77**, 7112–7116 (1980).
- Gilmore, R., Walter, P. & Blobel, G. Protein translocation across the endoplasmic reticulum. II. Isolation and characterization of the signal recognition particle receptor. *J. Cell Biol.* **95**, 470–477 (1982).
- Beckmann, R. *et al.* Alignment of conduits for the nascent polypeptide chain in the ribosome–Sec61 complex. *Science* **278**, 2123–2126 (1997).
- Menetret, J. F. *et al.* The structure of ribosome–channel complexes engaged in protein translocation. *Mol. Cell* **6**, 1219–1232 (2000).
- Beckmann, R. *et al.* Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107**, 361–372 (2001).
- Miller, J. D., Wilhelm, H., Gierach, L., Gilmore, R. & Walter, P. GTP binding and hydrolysis by the signal recognition particle during initiation of protein translocation. *Nature* **366**, 351–354 (1993).
- Miller, J. D., Bernstein, H. D. & Walter, P. Interaction of *E. coli* Ffh/4.5S ribonucleoprotein and FtsY mimics that of mammalian signal recognition particle and its receptor. *Nature* **367**, 657–659 (1994).
- Powers, T. & Walter, P. Reciprocal stimulation of GTP hydrolysis by two directly interacting GTPases. *Science* **269**, 1422–1424 (1995).
- Keenan, R. J., Freymann, D. M., Walter, P. & Stroud, R. M. Crystal structure of the signal sequence binding subunit of the signal recognition particle. *Cell* **94**, 181–191 (1998).
- Montoya, G., Svensson, C., Luijck, J. & Sinning, I. Crystal structure of the NG domain from the signal-recognition particle receptor FtsY. *Nature* **385**, 365–368 (1997).
- Freymann, D. M., Keenan, R. J., Stroud, R. M. & Walter, P. Structure of the conserved GTPase domain of the signal recognition particle. *Nature* **385**, 361–364 (1997).
- Ramirez, U. D. *et al.* Structural basis for mobility in the 1.1 Å crystal structure of the NG domain of *Thermus aquaticus* Ffh. *J. Mol. Biol.* **320**, 783–799 (2002).
- Vetter, I. R. & Wittinghofer, A. The guanine nucleotide-binding switch in three dimensions. *Science* **294**, 1299–1304 (2001).
- Montoya, G., Kaat, K., Moll, R., Schafer, G. & Sinning, I. The crystal structure of the conserved GTPase of SRP54 from the archaeon *Acidianus ambivalens* and its comparison with related structures suggests a model for the SRP–SRP receptor complex. *Struct. Fold. Des.* **8**, 515–525 (2000).
- Rapiejko, P. J. & Gilmore, R. Empty site forms of the SRP54 and SR α GTPases mediate targeting of ribosome–nascent chain complexes to the endoplasmic reticulum. *Cell* **89**, 703–713 (1997).
- Peluso, P., Shan, S. O., Nock, S., Herschlag, D. & Walter, P. Role of SRP RNA in the GTPase cycles of Ffh and FtsY. *Biochemistry* **40**, 15224–15233 (2001).
- Shan, S. & Walter, P. Induced nucleotide specificity in a GTPase. *Proc. Natl Acad. Sci. USA* **100**, 4480–4485 (2003).
- Padmanabhan, S. & Freymann, D. M. The conformation of bound GMPPNP suggests a mechanism for gating the active site of the SRP GTPase. *Structure* **9**, 859–867 (2001).

25. Smith, P. C. *et al.* ATP binding to the motor domain from an ABC transporter drives formation of a nucleotide sandwich dimer. *Mol. Cell* **10**, 139–149 (2002).
26. Schindelin, H., Kisker, C., Schlessman, J. L., Howard, J. B. & Rees, D. C. Structure of ADP × AIF4(–)-stabilized nitrogenase complex and its implications for signal transduction. *Nature* **387**, 370–376 (1997).
27. Seewald, M. J., Korner, C., Wittinghofer, A. & Vetter, I. R. RanGAP mediates GTP hydrolysis without an arginine finger. *Nature* **415**, 662–666 (2002).
28. Tesmer, J. J., Berman, D. M., Gilman, A. G. & Sprang, S. R. Structure of RGS4 bound to AIF4-activated Gi α1: Stabilization of the transition state for GTP hydrolysis. *Cell* **89**, 251–261 (1997).
29. Srinivasa, S. P., Watson, N., Overton, M. C. & Blumer, K. J. Mechanism of RGS4, a GTPase-activating protein for G-protein α subunits. *J. Biol. Chem.* **273**, 1529–1533 (1998).
30. Slep, K. C. *et al.* Structural determinants for regulation of phosphodiesterase by a G protein at 2.0 Å. *Nature* **409**, 1071–1077 (2001).
31. Focia, P. J., Shepotinovskaya, I. V., Seidler, J. A. & Freymann, D. M. Heterodimeric GTPase core of the SRP targeting complex. *Science* (in the press).
32. Otwinowski, Z. & Minor, W. Processing X-ray data in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1996).
33. Navaza, J. Implementation of molecular replacement in AMoRe. *Acta Crystallogr. D* **57**, 1367–1372 (2001).
34. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
35. Muller, K. *et al.* Moloc. *Bull. Soc. Chim. Belg.* **97**, 655–667 (1988).
36. Shepotinovskaya, I. V. & Freymann, D. M. Conformational change of the N-domain on formation of

the complex between the GTPase domains of *Thermus aquaticus* Ffh and FtsY. *Biochim. Biophys. Acta* **1597**, 107–114 (2002).

37. DeLano, W. L. *The PyMOL Molecular Graphics System* (<http://www.pymol.org/>) (2003).
38. Philippsen, A. DINO: *Visualizing Structural Biology* (<http://www.dino3d.org/>) (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Reyes for invaluable contributions to the initial FtsY mutant design and structure determination of *T. aquaticus* FtsY-GMPPNP, and R. Vale, H. Bourne and N. Bradshaw for comments on the manuscript. We acknowledge K. Slep and L. Rice for discussion and advice, and thank J. Holton and G. Meigg for support during data collection at the Advanced Light Source. D.F.S. was supported by a Burroughs-Wellcome Fund graduate fellowship. S.S. is supported by a Damon Runyan/Walter Winchell Cancer research fellowship. This work was supported by NIH grants to R.M.S. and P.W. P.W. is an Investigator of the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.M.S. (stroud@msg.ucsf.edu) or P.E.E. (pascal@msg.ucsf.edu). Coordinates for the FtsY–Ffh heterodimer complex bound to GMPPCP have been deposited in the Protein Data Bank under accession code 1RJ9.

An ultra-relativistic outflow from a neutron star accreting gas from a companion

Rob Fender¹, Kinwah Wu², Helen Johnston³, Tasso Tzioumis⁴, Peter Jonker⁵, Ralph Spencer⁶ & Michiel van der Klis¹

¹Astronomical Institute 'Anton Pannekoek', University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

²Mullard Space Science Laboratory, University College London, Holmbury St Mary, Surrey, RH5 6NT, UK

³School of Physics, University of Sydney, New South Wales 2006, Australia

⁴Australia Telescope National Facility, CSIRO, PO Box 76, Epping New South Wales 1710, Australia

⁵Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge, CB3 0HA, UK

⁶University of Manchester, Nuffield Radio Astronomy Laboratories, Jodrell Bank, Cheshire, SK11 9DL, UK

Collimated relativistic outflows—also known as jets—are amongst the most energetic phenomena in the Universe. They are associated with supermassive black holes in distant active galactic nuclei¹, accreting stellar-mass black holes and neutron stars in binary systems² and are believed to be responsible for γ -ray bursts³. The physics of these jets, however, remains

something of a mystery in that their bulk velocities, compositions and energetics remain poorly determined. Here we report the discovery of an ultra-relativistic outflow from a neutron star accreting gas within a binary stellar system. The velocity of the outflow is comparable to the fastest-moving flows observed from active galactic nuclei, and its strength is modulated by the rate of accretion of material onto the neutron star. Shocks are energized further downstream in the flow, which are themselves moving at mildly relativistic bulk velocities and are the sites of the observed synchrotron emission from the jet. We conclude that the generation of highly relativistic outflows does not require properties that are unique to black holes, such as an event horizon.

Circinus X-1 is a bright and highly variable X-ray source, containing a stellar-mass compact object accreting from a binary companion star. X-ray dips and outbursts with a period of 16.6 days are most readily interpreted as enhanced accretion during periastron passage of the accreting object in a significantly elliptical binary orbit^{4,5}. Observations of type I X-ray bursts indicate that the accreting object is a neutron star⁶. In the 1970s Cir X-1 was also associated with bright radio outbursts^{7,8}. Since then, there has been a decline in the strength of the radio counterpart, which has been found to reside within an extended radio nebula⁹. Structures on arcminute scales have been imaged within this nebula¹⁰, and more recently a one-sided jet on arcsecond scales, which aligns with the larger structures, has been discovered¹¹.

Bright radio flares associated with the production of a relativistic jet are now established to be a common property of accreting black

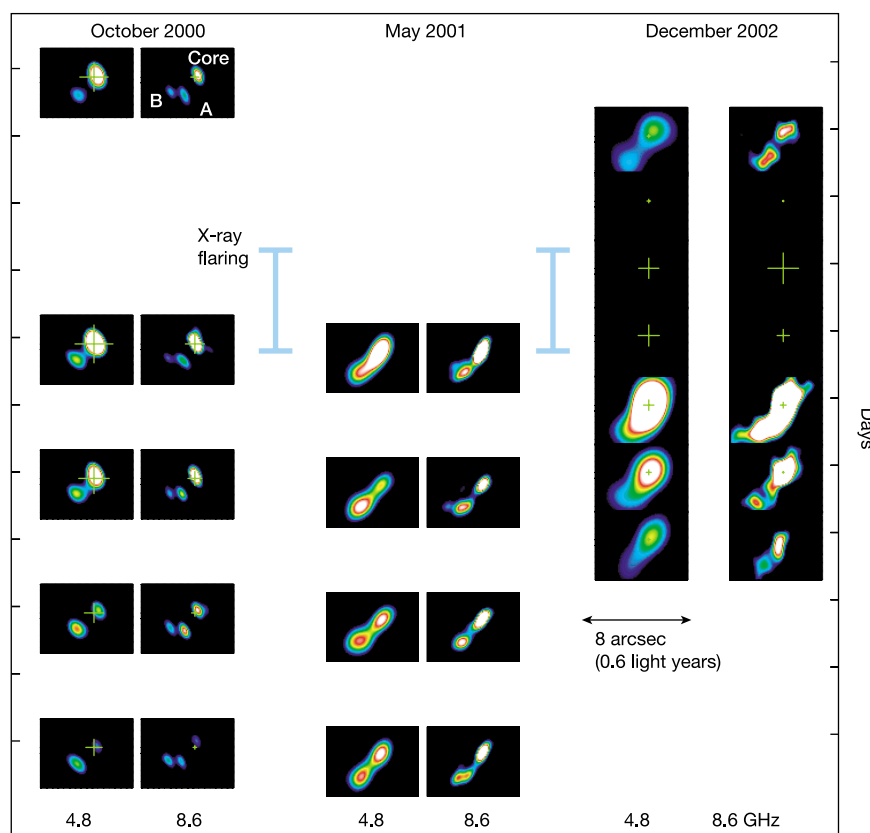


Figure 1 An ultrarelativistic outflow: sequences of radio observations of Circinus X-1 in October 2000, May 2001 and December 2002. At each epoch, observations were made simultaneously at 4.8 and 8.6 GHz. Tickmarks indicate time steps of one day; the blue bars indicate the time of the X-ray flaring as observed by the *Ross* X-ray Timing Explorer All-Sky monitor. In October 2000 and May 2001 the observations were spaced every two days; in December 2002 they were spaced daily. At each epoch the $u-v$ coverage of

the radio observations is identical for each image: maps in October 2000 and May 2001 are 'uniformly weighted' and those in December 2002 are 'naturally weighted'. The crosses indicate the location of the binary 'core' from ref. 11, and their size is proportional to the 'core' radio flux density. The observations reveal that after the X-ray flaring the extended radio structure brightens on timescales of days. The apparent velocities associated with this expansion are $\geq 15c$, indicating an underlying ultrarelativistic flow.

holes in outburst, but are less commonly associated with accreting neutron stars⁸. In several cases the radio emission has been resolved into outflowing components, sometimes either side of a stationary core, with bulk Lorentz factors (the Lorentz factor $\Gamma = (1 - \beta^2)^{-1/2}$, where velocity $v = \beta c$ and c is the speed of light) typically inferred to be in the range $2 \leq \Gamma \leq 5$ (ref. 2). These observations provide a generic qualitative picture in which electrons are accelerated to relativistic energies close to the accreting compact object, and then ejected from the system in a bipolar flow, from which we observe synchrotron emission until energy losses, primarily due to expansion of the plasma, reduce the emission below observable levels. However, very long baseline interferometric (VLBI) observations of the neutron-star binary Scorpius X-1 (refs 12, 13) have revealed moving radio components with mildly relativistic bulk velocities ($\beta \approx 0.5$) which are energized by a beam which is itself not directly detectable, but must have a Lorentz factor $\Gamma > 3$. Furthermore, it has been suggested that the observed synchrotron emission results from a moving shock¹⁴ (as discussed for active galactic nuclei (AGN) and γ -ray bursts), a possibility supported by recent observations of large-scale X-ray jets from a transient black hole X-ray binary, revealing *in situ* acceleration of electrons to TeV energies¹⁵.

At all frequencies and angular scales observed, Cir X-1 is associated with an extended radio structure. Careful inspection of Figs 1 and 2 reveals that X-ray flares are associated with a brightening of the 'core' of the extended radio structure. Previous comparison of Hubble Space Telescope and radio images¹¹ confirms the association of the radio 'core' with the optical counterpart of the binary, and so we associate the extended emission or

knots with an outflow from the system. The radio spectral indices ($-1.1 \leq \alpha \leq -0.7$) indicate optically thin synchrotron emission.

We shall focus initially on the observations of October 2000 (left panels of Figs 1 and 2), which were sampled every second day. In the 'quiescent' observation before the X-ray flaring the radio source at 4.8 GHz is clearly associated with a bright knot near to the core (best radio position from ref. 11), plus a weaker extended structure. At 8.6 GHz the extended emission is resolved into two components, knots A and B, which are not co-aligned with the core. After the X-ray flaring, the core brightened significantly within one day, and within three days the extended emission brightened, peaking within five days of the outburst. Subsequently the extended emission faded. In May 2001 (middle panels of Fig. 1) we performed a similar set of observations with two-day sampling and saw a similar effect. After the X-ray flaring the extended emission brightens at 4.8 GHz within two or three days, and there seems to be a 'bridge' of radio emission from the core to the knot immediately after the flare; in addition, in the difference maps (Fig. 2) there is some indication of a counterjet. In December 2002 we made a sequence of observations with daily sampling. This time, after the X-ray flare the core brightened strongly and a resolved bridge of radio emission appeared along the apparently curved axis of the extended emission.

At all epochs we clearly see evidence for the transmission of energy from the binary core to the extended radio emission following an outburst. We investigated the apparent velocity this signal might correspond to. For a proper motion μ , the apparent

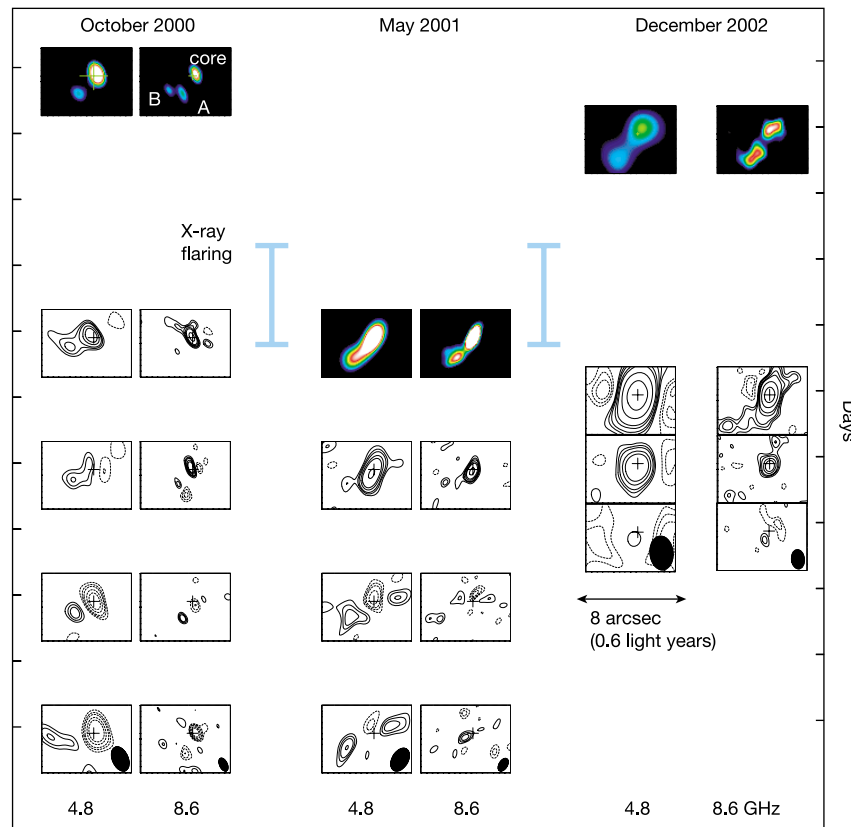


Figure 2 The same data as in Fig. 1, but presented in the form of 'difference' maps. At each of the three epochs the top (reference) image at each frequency is the same as that presented in Fig. 1. The contour images below it correspond to the maps from the corresponding frames in Fig. 1 with the reference image subtracted, to clearly illustrate changes in the radio structure. Solid lines indicate positive residuals, dashed lines indicate

negative residuals; contour levels have been chosen to indicate the level of noise in each image. At each epoch, the strongest variability in the difference maps is clearly occurring in the region of the jet or knots. Solid ellipses in the lower right corners of the bottom panels indicate the size and shape of the synthesised beam at each epoch and frequency.

velocity expressed as a fraction of the speed of light $\beta_{\text{app}} = v_{\text{app}}/c$, for the source at a distance d is:

$$\beta_{\text{app}} \approx \left(\frac{d}{\text{kpc}} \right) \left(\frac{\mu}{170 \text{ mas d}^{-1}} \right)$$

All the maps show the brightening of the extended components within days of the X-ray flare. At all epochs, the extended emission at 4.8 GHz, at a distance of 2.2–2.5 arcsec from the core, brightened within three to five days. From these numbers we can estimate a conservative lower limit to the proper motion associated with the energizing beam of 400 mas d^{-1} . The most likely distance to Cir X-1 is $\sim 6.5 \text{ kpc}$ (ref. 10). From this we can establish that the apparent signal propagation velocity between the core and knot is $\beta_{\text{app}} > 15$. For an intrinsic velocity β at an angle θ to the line of sight:

$$\beta_{\text{app}} = \frac{\beta \sin \theta}{(1 - \beta \cos \theta)}$$

which has a maximum value of $\beta_{\text{app}} = \Gamma\beta$ when $\beta = \cos \theta$. Because the observations indicate $\beta_{\text{app}} > 15$, this implies $\Gamma\beta > 15$, for which the minimum velocity solution is $\Gamma > 15$ (that is, $\beta > 0.998$). This is by far the most relativistic flow observed to date within our Galaxy. Furthermore, for any value of β , solutions may only be obtained for angles to the line of sight of $\theta < 5^\circ$, indicating that the signal responsible for the brightening of the knots must propagate at an angle very close to the line of sight. The images, in particular at 8.6 GHz, give a clear indication of bending of the jet; by comparison, with AGN we can also infer a small angle to the line of sight, in which case projection effects increase the apparent bend angle.

This lower limit to the Lorentz factor may correspond to the bulk velocity of the flow between the core and the jets, or of a pattern (shock) propagating along this flow. In the latter case, in the approximation of a one-dimensional relativistic shock propagating along a relativistic fluid, the bulk (Γ_{bulk}) and shock (Γ_{shock}) Lorentz factors are related by $\Gamma_{\text{bulk}} = \Gamma_{\text{shock}}/\sqrt{2}$ in the limit (appropriate here) of $\beta_{\text{shock}} \rightarrow 1$ (see ref. 16 for example). Thus our lower limit to the signal Lorentz factor of $\Gamma \geq 15$ may place a lower limit on the bulk Lorentz factor of the flow of $\Gamma_{\text{bulk}} \geq 10$.

Further clues to the nature of the sites of emission are provided by linear polarization maps, to be presented elsewhere, which reveal a magnetic field in the knots which is parallel to the jet axis in the core and perpendicular to the jet axis at the knots. This supports an interpretation of the extended emission as synchrotron emission from shocks formed at a longitudinal compression of the jet flow, as in AGN¹⁷. Furthermore, it is clear that there is secular evolution of the mean radio structure between the epochs, with a general trend for the propagation of components away from the binary core. The projected velocity of these components is in the order of $0.01c$.

These observations reveal that underlying observed radio jets of the accreting neutron star Cir X-1 is a flow with $\Gamma_{\text{bulk}} \geq 10$. In another neutron star binary, Sco X-1, the underlying physics may be similar^{12,13}. Sco X-1 is the prototype of the class of ‘Z sources’, representing the six most-luminous neutron-star accretors in our Galaxy, and it has further been suggested that Cir X-1 itself shares some of the properties of this class¹⁸. Thus it seems plausible that production of such ultrarelativistic jets may be a generic feature of accretion onto neutron stars at or near to the Eddington limit—this is at odds with the popular suggestion that the velocities of jets from neutron stars (and indeed all accreting objects) would be limited to the surface escape speed, in this case $\sim 0.3c$ (ref. 19).

The minimum energy associated with brightening of the extended emission is $E_{\text{min}} \approx 10^{40} \text{ erg}$, for a source volume of radius one light day (the estimated energization timescale); the associated power input rate is $P_j \geq 10^{35} \text{ erg s}^{-1}$. If the unseen energizing signal were isotropic—for example, a spherical blast wave from the

binary—this would correspond to a total power output rate of $\geq 7 \times 10^{39} \text{ erg s}^{-1}$, more than an order of magnitude above the Eddington limit for a neutron star. Using these arguments we interpret the signal as rising in a collimated ultrarelativistic flow.

The velocities of jets from black holes in our Galaxy are not well determined; in particular it is almost impossible to place an upper limit on the Lorentz factors of the observed motions²⁰. In the case of AGN it has been argued that the bulk Lorentz factor of motion is $\Gamma_{\text{bulk}} \approx 10$ (ref. 21), although in a small fraction of blazars apparent velocities $\beta \geq 30 h^{-1}$ (where the Hubble constant $H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}$) have been measured on VLBI scales²², with the same implications for the underlying bulk velocity as derived here for Cir X-1. Thus these observations demonstrate that the study of jets from galactic X-ray binaries, whether containing neutron-star or black-hole accretors, may be directly applied to our study of jets from the most powerful engines in the Universe, the AGN. Furthermore, the discovery of such ultrarelativistic flows associated with a neutron star points us clearly to the common features of neutron-star and black-hole accretion, such as the accretion flow, and not properties unique to black holes, such as an event horizon or ergosphere, as the sites of the key physics in jet formation. $\square$

Received 8 July; accepted 8 October 2003; doi:10.1038/nature02137.

- Ostrowski, M., Sikora, M., Madejski, G. & Begelman, M. (eds) *Relativistic Jets in AGNs* (Towarzystwo Salezjańskiego, Krakow, 1997).
- Mirabel, I. F. & Rodríguez, L. F. Sources of relativistic jets in the Galaxy. *Annu. Rev. Astron. Astrophys.* **37**, 409–443 (1999).
- Sari, R., Piran, T. & Halpern, J. P. Jets in gamma-ray bursts. *Astrophys. J.* **519**, L17–L20 (1999).
- Kaluźniński, L. J., Holt, S. S., Voldt, E. A. & Serlemitsos, P. J. Evidence for a 16.6 day period from Circinus X-1. *Astrophys. J.* **208**, L71–L75 (1976).
- Johnston, H. M., Fender, R. & Wu, K. High-resolution optical and infrared spectroscopic observations of Cir X-1. *Mon. Not. R. Astron. Soc.* **308**, 415–423 (1999).
- Tennant, A. F., Fabian, A. C. & Shafer, R. A. Observation of type I X-ray bursts from Cir X-1. *Mon. Not. R. Astron. Soc.* **221**, P27–P31 (1986).
- Haynes, R. F. *et al.* Simultaneous radio and optical observations of Cir X-1. *Mon. Not. R. Astron. Soc.* **185**, 661–671 (1978).
- Fender, R. P. & Kuulkers, E. On the peak radio and X-ray emission from neutron star and black hole candidate X-ray transients. *Mon. Not. R. Astron. Soc.* **324**, 923–930 (2001).
- Haynes, R. F. *et al.* A radio nebula associated with Circinus X-1. *Nature* **324**, 233–235 (1986).
- Stewart, R. T., Caswell, J. L., Haynes, R. F. & Nelson, G. J. Circinus X-1—A runaway binary with curved radio jets. *Mon. Not. R. Astron. Soc.* **261**, 593–598 (1993).
- Fender, R. P. *et al.* An asymmetric arcsecond radio jet from Circinus X-1. *Astrophys. J.* **506**, L121–L125 (1998).
- Fomalont, E. B., Geldzahler, B. J. & Bradshaw, C. F. Scorpius X-1: Energy transfer from the core to the radio lobes. *Astrophys. J.* **553**, L27–L30 (2001).
- Fomalont, E. B., Geldzahler, B. J. & Bradshaw, C. F. Scorpius X-1: The Evolution and nature of the twin compact radio lobes. *Astrophys. J.* **558**, 283–301 (2001).
- Kaiser, C. R., Sunyaev, R. & Spruit, H. C. Internal shock model for microquasars. *Astron. Astrophys.* **356**, 975–988 (2000).
- Corbel, S. *et al.* Large-scale, decelerating, relativistic X-ray jets from the microquasar XTE J1550–564. *Science* **298**, 196–199 (2002).
- Krolik, J. H. *Active Galactic Nuclei* (Princeton Series in Astrophysics, Princeton Univ. Press, Princeton, USA, 1999).
- Saikia, D. J. & Salter, C. J. Polarization properties of extragalactic radio sources. *Annu. Rev. Astron. Astrophys.* **26**, 93–144 (1988).
- Shirey, R. E., Bradt, H. V. & Levine, A. M. The complete ‘Z’ track of Circinus X-1. *Astrophys. J.* **517**, 472–487 (1999).
- Livio, M. Astrophysical jets: a phenomenological examination of acceleration and collimation. *Phys. Rep.* **311**, 225–245 (1999).
- Fender, R. P. Uses and limitations of relativistic jet proper motions: Lessons from galactic microquasars. *Mon. Not. R. Astron. Soc.* **340**, 1353–1358 (2003).
- Ghisellini, G., Padovani, P., Celotti, A. & Maraschi, L. Relativistic bulk motion in active galactic nuclei. *Astrophys. J.* **407**, 65–82 (1993).
- Jorstad, S. G. *et al.* Multiepoch Very Long Baseline Array observations of EGRET-detected quasars and BL Lacertae objects: Superluminal motion of gamma-ray bright blazars. *Astrophys. J. Supp. Ser.* **134**, 181–240 (2001).

Acknowledgements The Australia Telescope is funded by the Commonwealth of Australia for operation as a national facility managed by CSIRO.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.F. (rpf@science.uva.nl).

Probable observation of a supersolid helium phase

E. Kim & M. H. W. Chan

Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

When liquid ^4He is cooled below 2.176 K, it undergoes a phase transition—Bose–Einstein condensation—and becomes a superfluid with zero viscosity¹. Once in such a state, it can flow without dissipation even through pores of atomic dimensions. Although it is intuitive to associate superflow only with the liquid phase², it has been proposed theoretically^{3–5} that superflow can also occur in the solid phase of ^4He . Owing to quantum mechanical fluctuations, delocalized vacancies and defects are expected to be present in crystalline solid ^4He , even in the limit of zero temperature. These zero-point vacancies can in principle allow the appearance of superfluidity in the solid^{3,4}. However, in spite of many attempts⁶, such a ‘supersolid’ phase has yet to be observed in bulk solid ^4He . Here we report torsional oscillator measurements on solid helium confined in a porous medium, a configuration that is likely to be more heavily populated with vacancies than bulk helium. We find an abrupt drop in the rotational inertia⁵ of the confined solid below a certain critical temperature. The most likely interpretation of the inertia drop is entry into the supersolid phase. If confirmed, our results show that all three states of matter—gas⁷, liquid¹ and solid—can undergo Bose–Einstein condensation.

The most direct experiment searching for the supersolid phase in bulk solid ^4He (performed by Bishop, Paalanen and Reppy⁸) also used the torsional oscillator technique. The resonant period of the

high- Q oscillator shown in Fig. 1 is given by $2\pi\sqrt{I/G}$, where I is the moment of inertia of the torsion bob, which contains helium, and G is the torsional spring constant of the Be–Cu torsion rod. A small hole drilled through the centre of the torsion rod allows the introduction of helium into the torsion bob. The oscillator is driven and maintained at resonance by a pair of electrodes. The onset of superfluidity in the helium inside the torsion bob decreases I , and hence decreases the resonant period. Bishop *et al.*⁸ made measurements of solid helium from 25 to 48 bar, and concluded that if there is a supersolid state, then either the supersolid fraction (the fraction of ^4He atoms participating in superflow) is less than 5×10^{-6} or the critical velocity is less than $5 \mu\text{m s}^{-1}$. (The critical velocity is the maximum velocity of superflow without any detectable dissipation.)

In contrast to the results of Bishop *et al.*⁸, our torsional oscillator measurements on solid helium grown inside a porous Vycor glass disk show a decrease in the resonant period, characteristic of entry into a supersolid state. The continuous pore space, constituting 30% of the total volume in Vycor, appears under the transmission electron microscope as a network of randomly and multiply interconnected cylindrical channels of about 7 nm diameter and 30 nm length⁹. There have been a number of experiments^{10–15}, including a torsional oscillator measurement by Brewer and collaborators^{10,11}, studying the solidification of ^4He inside Vycor glass. ^4He remains liquid down to temperature $T = 0$ K, unless a substantial pressure is applied to the sample. Below 1.3 K, this freezing pressure is essentially constant at 25 bar. Inside Vycor glass, however, a pressure close to 40 bar is required for solidification^{10–15}. At low temperature in the presence of ^4He vapour, an amorphous surface film is adsorbed on the walls of the pores by the van der Waals potential. Because of lattice mismatch, this surface film is not favourable for the nucleation and continued growth of solid as the pressure is increased and brought towards the bulk freezing pressure. This means that freezing is initiated from the liquid in the centre of the pore by homogeneous nucleation of crystallites of radius limited by the pore size. The overpressure required to seed a crystallite of

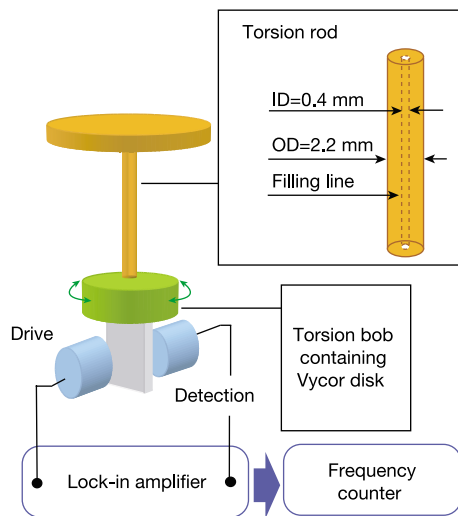


Figure 1 Torsional oscillator used in this experiment. The design of the oscillator follows those used by Reppy and collaborators¹⁸. The Vycor glass disk has a diameter of 15 mm and a thickness of 4 mm. The cylindrical drive and detection electrodes are aligned off-centre from, and are capacitively coupled to, the central electrode attached to the torsion bob. The signal from the detection electrode (proportional to the amplitude) is sent to the lock-in amplifier through a current preamplifier. The lock-in provides a driving voltage, which controls the amplitude of oscillation, to complete the phase-locked loop and keep the oscillator in resonance. The mechanical Q of the oscillator is 10^6 at low temperature, allowing the determination of the resonant period to a precision of 0.2 ns. The resonant period is 967,640 ns when the Vycor disk is empty, and is 971,900 ns near 0.2 K when pressurized with solid ^4He at 62 bar. Measurements were also made with a dummy torsional cell with the Vycor glass disk replaced by a solid brass disk.

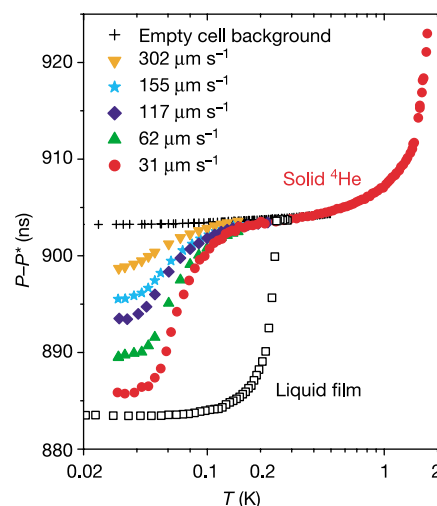


Figure 2 Resonant period as function of temperature of solid ^4He in Vycor glass. The resonant period for different oscillation amplitudes—and hence different velocities of the rim of the Vycor disk, v_{rim} —is shown. A drop in the period (ΔP), signifying the transition into the supersolid phase, is seen below 175 mK. Although the magnitude of ΔP depends strongly on the rim velocity, no such dependence of the period is seen above the transition temperature. For comparison, the empty (without helium) cell period, and the period of an atomically thin liquid film adsorbed on the walls of the internal pore space of Vycor, are also shown. The film measurement, showing a superfluid transition at 250 mK, is carried out with the same torsion cell. For easy comparison, 4,260 ns is added to the empty cell data and 3,290 ns to the film data. The ordinate shows $P - P^*$, the difference of the actual period P and $P^* = 971,000$ ns.

radius r in this geometric model of freezing is predicted to be proportional to the interfacial tension between the liquid and solid phases and inversely proportional to r . The large (15 bar) overpressure observed for solidification of ^4He in Vycor reflects its small pore diameter^{10–15}. Solid ^4He grown inside Vycor with a tortuous porous structure and small characteristic pore diameter is likely to be heavily populated with vacancies.

The solid sample is grown and its density kept constant via the standard blocked capillary method^{10–15}. We found that by cooling (from 3 K) a liquid sample of 75 bar in the torsion cell, the resultant pressure of the solid sample below 1 K is typically 62 ± 2 bar. The pressure is determined by an *in situ* strain gauge attached directly on the outside of the torsion cell. We have intentionally chosen this high pressure so that we can be sure that our helium sample in Vycor glass is deep in the solid phase. Figure 2 shows the resonant period of our Vycor disk torsional oscillator as a function of temperature. As the mechanical Q of the oscillator is 10^6 , and the resonant period is of the order of 1 ms, the equilibration time for a new resonant period reading due to any change in the experimental condition—for example, a change in temperature—is expected and observed to be of the order of 1,000 s or 15 min.

Below 80 mK, however, the equilibration time of the resonant period depends on whether the measurement is taken during warming or cooling of the torsional cell. In a warming scan, the equilibration time is the same as that at higher temperatures; but in a cooling scan, this time lengthens noticeably with decreasing temperature and exceeds 60 min below 40 mK. Our period data are obtained during warming scans after waiting for complete equilibration at the lowest temperature. The resonant period was measured with different amplitudes of oscillations. The amplitude, proportional to the linear velocity of the rim of the Vycor disk, can be calculated from the a.c. voltage induced on the detection electrode. The resonant period above 0.2 K does not depend on the rim velocity, and for temperatures above 0.5 K it is similar to that found by Brewer and collaborators^{10,11}. These authors did not extend their measurement below 0.5 K. The central result of our experiment is the additional drop in the period (ΔP) that begins at 175 mK. This drop is consistent with the confined solid entering the supersolid phase. ΔP is strongly attenuated with increasing amplitude of oscillation (that is, higher rim velocity of the Vycor). Figure 3 shows ΔP in the low-temperature limit as a function of the rim velocity. The transition temperature at 175 mK, however, does not appear to depend on the rim velocity.

For a disk oscillating at a fixed angular velocity, the local linear velocity ranges from zero up to a maximum rim velocity. Supersolid decoupling occurs for solid embedded in the region where the local

velocity is smaller than the critical velocity. The temperature dependence of ΔP shown in Fig. 2 is probably a reflection of the velocity profile of the Vycor disk modulated by the temperature-dependent critical velocity of the supersolid. Although we do not know the exact functional form, the critical velocity is zero at the transition temperature (175 mK), increases with decreasing temperature, and saturates near $300 \mu\text{m s}^{-1}$ in the low-temperature (30 mK) limit. As a comparison, we also show in Fig. 2 the superfluid response of an atomically thin liquid film adsorbed in Vycor with zero-temperature ΔP of the same order as that of the solid sample. The measurements were carried out with the same torsional cell. In addition to the different temperature dependence, ΔP of the film as shown in Fig. 3 is independent of the rim velocity. This is not surprising, as the critical velocity of an atomically thin liquid ^4He film adsorbed in Vycor exceeds 200 mm s^{-1} (ref. 16), which is a factor of 700 higher than the maximum rim velocity of $300 \mu\text{m s}^{-1}$ used in this experiment.

Superfluid helium adsorbed inside an oscillating porous disk must execute a tortuous path of potential flow defined by pore structure¹⁷. As a result, a fraction, χ , of the superfluid moment of inertia, I_S , remains effectively locked to the porous disk and the observed period drop is proportional to the unlocked portion, $(1 - \chi)I_S$. If we make the assumption that the χ factor of supersolid in Vycor is zero, then the ΔP_0 of 22 ns shown in Fig. 3 is a direct measure of the zero-temperature and low-velocity supersolid

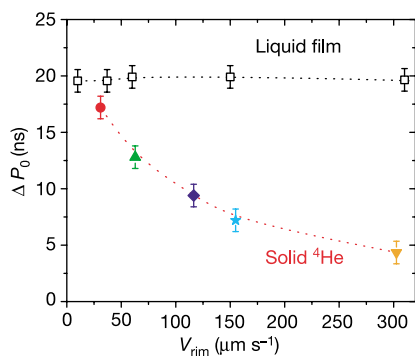


Figure 3 ΔP at the low temperature limit, ΔP_0 , as a function of rim velocity of the Vycor disk. ΔP_0 values are deduced by subtracting the measured periods at 30 mK from the (shifted) empty cell period, as shown in Fig. 2. Whereas ΔP_0 of the liquid film is independent of rim velocity, it is strongly attenuated with higher rim velocity in solid ^4He . This plot shows that the critical velocity of supersolid at 30 mK is of the order of $300 \mu\text{m s}^{-1}$. It also shows that ΔP_0 extrapolates to 22 ns in the limit of low rim velocity.

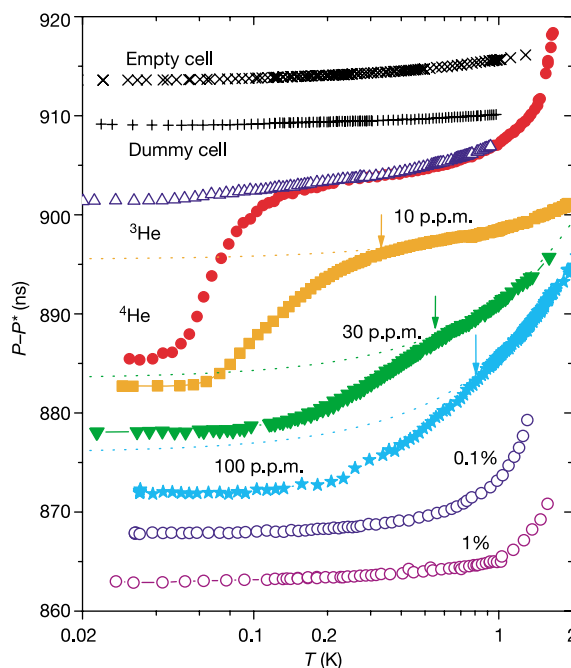


Figure 4 Resonant periods as a function of temperature for a variety of solid helium samples. The period scale shown corresponds to that for solid ^4He . As in Fig. 2, the ordinate shows $P - P^*$, the difference of the actual period P and $P^* = 971,000$ ns. The period data for other samples are shifted for clarity and easy comparisons. All measurements were made with the rim velocity of the Vycor disk near $30 \mu\text{m s}^{-1}$. The plots show that the period drop effect is not related to the stiffening of bulk solid helium in the torsion rod. The effect is not seen in pure ^3He , and is not seen in solid mixtures with ^3He concentration exceeding 0.1%. The resonant periods in these samples are independent of the rim velocity of the Vycor disk. A period drop is found for mixtures with 10, 30 and 100 p.p.m. of ^3He . As in pure ^4He , the size of the drop in these samples with low ^3He concentrations is rim-velocity dependent. The dotted lines extrapolated smoothly from high temperature are the expected background period in the absence of period drops. The vertical arrows mark the transition temperatures of these samples. The ^3He concentrations listed are the average concentration of the solid inside the Vycor and in the capillary leading to the cell. The actual ^3He concentration inside the Vycor, particularly for low-concentration samples, could be quite different from the listed values.

moment of inertia. The resonant period increases by 4,260 ns when the porous Vycor glass disk in the torsion bob is filled with solid ^4He at 62 bar, therefore the supersolid fraction—the fraction of ^4He atoms that participate in superflow—is 5 parts in 10^3 . On the other hand, if we assume that the χ factor of supersolid is 0.8, the same as that of superfluid in Vycor¹⁸, then the supersolid fraction is 25 parts in 10^3 . If we assume that the ^4He atoms in the supersolid fraction are uniformly distributed in the pore space and if we neglect interaction between these atoms, then we can use the ideal Bose gas theory to calculate the Bose–Einstein condensation temperature¹⁹. The calculated transition temperatures are 120 mK and 350 mK, respectively, for supersolid fractions of 5 and 25 parts in 10^3 , bracketing the observed transition temperature of 175 mK.

In order to rule out non-supersolid mechanisms for the observed effect, we did a number of control experiments. We made measurements with the same torsional cell with pure solid ^3He and with solid ^4He diluted with 10, 30, 100, 1,000 and 10,000 p.p.m. of ^3He , all pressurized with the same procedures as that for pure ^4He , and resulting in a final pressure between 60 and 65 bar. The results of these measurements are shown in Fig. 4. The observed ΔP seen in pure ^4He is not seen for solid ^3He , and is also not seen for solid mixtures with ^3He concentrations exceeding 0.1%. The fact that the effect is not present in ^3He , a Fermi system, is reassuring. It is also reasonable that the addition of enough ^3He atoms is effective in quenching the supersolid phase. The behaviour found in samples with even lower ^3He concentrations is very intriguing. Besides reducing the magnitude of ΔP , the introduction of the minute amount of ^3He also broadens the transition and increases the transition temperature.

Figure 4 also shows measurements made with a dummy torsional oscillator consisting of a torsion rod that is identical to the normal torsional cell but with a torsion bob containing a solid brass disk rather than a Vycor glass disk. During measurements, the hole in the torsion rod is pressurized with solid ^4He (using the exact same procedure as that for the Vycor torsional cell) to a final pressure of 62 bar. The resonant period is temperature independent, showing no decrease at low temperature, similar to that of the empty cell. This means the effect that we have seen with pure solid ^4He and with ^4He diluted with small ^3He impurities occurs inside the torsion cell, and is not related to changes occurring in the bulk solid helium inside the torsion rod. Owing to the freezing of dislocations, solid helium inside the torsion rod stiffens at low temperature. This stiffening can contribute to the torsional spring constant of the torsion rod²⁰ and lower the resonant period. We have, however, taken the precaution of using an especially thick (2.2 mm diameter) torsion rod, so that this effect can be ignored. This assumption is confirmed by our measurements with the dummy cell.

It could be argued that solidification inside the Vycor glass is never complete, and that there is a persistent thin liquid film even at 62 bar, 20 bar above the reported solidification pressure. But there are observations that are not consistent with this picture. The temperature dependence of the supersolid is distinctly different from that of a liquid film, and the critical velocity observed for the solid is at least 700 times smaller than that in a film¹⁶. Figure 4 shows that the introduction of 0.1% of ^3He into solid ^4He completely eliminates the observed drop in the period. This is in strong contrast to the findings in liquid films. The addition of ^3He into a liquid ^4He film smoothly decreases the superfluid transition temperature, and a very high concentration of ^3He is needed to completely quench the transition²¹. The transition temperature of a pure ^4He film at 150 mK was found to decrease to below 20 mK only when the amount of ^3He exceeds 20% of all the ^4He in the adsorbed film, including that in the amorphous solid layer²¹.

In conclusion, the most reasonable interpretation of the observed period drop is that it is a signature of transition into the supersolid state. The microscopic origin of this effect is not understood. We have noted that the solid ^4He grown in the Vycor pores is heavily

populated with vacancies and defects. It seems that these vacancies may be responsible for enhancing Bose–Einstein condensation of the confined solid ^4He into the supersolid phase. □

Received 23 September; accepted 14 November 2003; doi:10.1038/nature02220.

1. Kapitza, P. Viscosity of liquid helium below the λ -point. *Nature* **141**, 74 (1938).
2. Penrose, O. & Onsager, L. Bose–Einstein condensation and liquid helium. *Phys. Rev.* **104**, 576–584 (1956).
3. Andreev, A. F. & Lifshitz, I. M. Quantum theory of defects in crystals. *Sov. Phys. JETP* **29**, 1107–1113 (1969).
4. Chester, G. V. Speculations on Bose–Einstein condensation and quantum crystals. *Phys. Rev. A* **2**, 256–258 (1970).
5. Leggett, A. J. Can a solid be “superfluid”? *Phys. Rev. Lett.* **25**, 1543–1546 (1970).
6. Meisel, M. W. Supersolid ^4He —An overview of past searches and future possibilities. *Physica B* **178**, 121–128 (1992).
7. Anderson, M. H., Ensher, J. R., Matthews, M. R., Wieman, C. E. & Cornell, E. A. Observation of Bose–Einstein condensation in a dilute atomic vapour. *Science* **269**, 198–201 (1995).
8. Bishop, D. J., Paalanen, M. A. & Reppy, J. D. Search for superfluidity in hcp ^4He . *Phys. Rev. B* **24**, 2844–2845 (1981).
9. Levitz, P., Ehret, G., Sinha, S. K. & Drake, J. M. Porous Vycor glass—the micro structure as probed by electron-microscopy, direct energy-transfer, small-angle scattering, and molecular adsorption. *J. Chem. Phys.* **95**, 6151–6161 (1991).
10. Brewer, D. F., Cao, L., Girit, C. & Reppy, J. D. ^4He transition in a restricted geometry below and above the bulk solidification pressure. *Physica B* **107**, 583–584 (1981).
11. Cao, L., Brewer, D. F., Girit, C., Smith, E. N. & Reppy, J. D. Flow and torsional oscillator measurements on liquid helium in restricted geometries under pressure. *Phys. Rev. B* **33**, 106–117 (1986).
12. Beamish, J. R., Hikata, A., Tell, L. & Elbaum, C. Solidification and superfluidity of ^4He in porous Vycor glass. *Phys. Rev. Lett.* **50**, 425–428 (1983).
13. Molz, E. B. & Beamish, J. R. Freezing and melting of helium in different porous media. *J. Low-Temp. Phys.* **101**, 1055–1077 (1995).
14. Adams, E. D., Uhlig, K., Tang, Y. H. & Haas, G. E. Solidification and superfluidity of ^4He in confined geometries. *Phys. Rev. Lett.* **52**, 2249–2252 (1984).
15. Bittner, D. N. & Adams, E. D. Solidification of helium in confined geometries. *J. Low-Temp. Phys.* **97**, 519–535 (1994).
16. Chan, M. H. W., Yanof, A. W. & Reppy, J. D. Superfluidity of thin ^4He films. *Phys. Rev. Lett.* **32**, 1347–1350 (1974).
17. Mehl, J. B. & Zimmermann, W. Jr Flow of superfluid helium in a porous medium. *Phys. Rev.* **167**, 214–229 (1968).
18. Berthold, J. E., Bishop, D. J. & Reppy, J. D. Superfluid transition of ^4He films adsorbed on porous Vycor glass. *Phys. Rev. Lett.* **39**, 348–352 (1977).
19. Huang, K. *Statistical Mechanics* 2nd edn, 293 (Wiley & Son, New York, 1967).
20. Palanen, M. A., Bishop, D. J. & Dail, H. W. Dislocation motion in hcp ^4He . *Phys. Rev. Lett.* **46**, 664–667 (1981).
21. Csáthy, G. A. & Chan, M. H. W. Effect of ^3He on submonolayer superfluidity. *Phys. Rev. Lett.* **87**, 045301 (2001).

Acknowledgements We acknowledge discussions with J. Banavar, J. Beamish, V. Crespi, J. Goodkind, J. Jain, A. Leggett and J. Reppy. This work was supported by the Condensed Matter Physics Program of the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.H.W.C. (chan@phys.psu.edu).

Partial order in the non-Fermi-liquid phase of MnSi

C. Pfleiderer¹, D. Reznik^{3,4}, L. Pintschovius³, H. v. Löhneysen^{1,3}, M. Garst² & A. Rosch²

¹Physikalisches Institut, ²Institut für Theorie der Kondensierten Materie, Universität Karlsruhe, D-76128 Karlsruhe, Germany

³Forschungszentrum Karlsruhe, Institut für Festkörperphysik, D-76021 Karlsruhe, Germany

⁴Laboratoire Léon Brillouin, CEA Saclay, F-91191 Gif-sur-Yvette Cedex, France

Only a few metallic phases have been identified in pure crystalline materials. These include normal, ferromagnetic and anti-ferromagnetic metals, systems with spin and charge density wave order, and superconductors. Fermi-liquid theory provides a basis for the description of all of these phases. It has been suggested

that non-Fermi-liquid phases of metals may exist in some heavy-fermion compounds^{1,2} and oxide materials^{3–6}, but the discovery of a characteristic microscopic signature of such phases presents a major challenge. The transition-metal compound MnSi above a certain pressure ($p_c = 14.6$ kbar) provides what may be the cleanest example of an extended non-Fermi-liquid phase in a three-dimensional metal^{7–9}. The bulk properties of MnSi suggest that long-range magnetic order is suppressed at p_c (refs 7–12). Here we report neutron diffraction measurements of MnSi, revealing that sizeable quasi-static magnetic moments survive far into the non-Fermi-liquid phase. These moments are organized in an unusual pattern with partial long-range order. Our observation supports the existence of novel metallic phases with partial ordering of the conduction electrons (reminiscent of liquid crystals), as proposed for the high-temperature superconductors^{4–6} and heavy-fermion compounds¹³.

The identification of novel metallic phases, as opposed to complicated crossover phenomena, has become controversial for a number of reasons. (1) The behaviour is observed in entirely new classes of materials for which little is known in general. (2) The role of dimensionality is unclear. (3) Metallurgical complexities exist, such as chemical segregation and defects. (4) The properties are extremely sensitive to a careful fine-tuning of the experimental conditions. (5) Characteristic energy scales are all of similar magnitude, and cannot be distinguished properly; examples are fluctuations of spin, charge or superconducting order in the high- T_c copper oxides.

The transition-metal compound MnSi is devoid of these complexities. (1) It is a very well-known system, and is perhaps the most extensively studied itinerant-electron magnet apart from iron, cobalt, nickel and chromium. (2) A large body of thermodynamic and microscopic data^{10–12,14–20} establish the ground state below the magnetic ordering temperature $T_c = 29.5$ K as a three-dimensional weakly spin-polarized Fermi liquid *par excellence*, with the possible exception of recent high-frequency optical conductivity experiments²¹. (3) As a congruently melting compound, MnSi can be produced at high purity and high crystalline perfection in the cubic B20 structure. (4) At the first-order magnetic quantum phase transition at $p_c = 14.6$ kbar, a discontinuous change from Fermi-liquid behaviour to an extended non-Fermi-liquid (NFL) phase has been inferred from the resistivity^{7–9}, which is proportional to $T^{1.5}$ for almost three decades in temperature T from around 6 K down to a few mK and for a large pressure range above p_c . This suggests that the novel behaviour does not require careful fine-tuning. (5) Three well-separated energy and length scales can be distinguished in MnSi at all temperatures, pressures and magnetic fields as follows. First, MnSi has a strong tendency to itinerant ferromagnetism on length scales of a few lattice constants, $a = 4.56$ Å, with an ordered moment of about $0.4 \mu_B$ per formula unit (where μ_B is the Bohr magneton). Second, as the B20 structure lacks an inversion symmetry, weak spin-orbit interactions assume a Dzyaloshinsky-Moriya (DM) form, $\int \mathbf{S} \cdot (\nabla \times \mathbf{S}) d\mathbf{r}$, which (being linear in momentum) destabilizes the uniform ferromagnetic order and introduces a well-understood^{22,23} helical modulation of a long wavelength 175 Å at ambient pressure. (Here $\mathbf{S}(\mathbf{r})$ is the space dependent magnetization.) Third, in the ordered phase, further spin-orbit interactions induced by the cubic crystalline electric fields lock the direction of the spiral to $\mathbf{Q} = \langle 111 \rangle$, where $\mathbf{S} \perp \mathbf{Q}$ (refs 19, 20, 22, 23). Typical sizes of magnetic domains in the ordered state are 10^4 Å (ref. 20). The locking of the direction of the helix hence represents the weakest scale.

For our study the helical modulation at ambient pressure proves to be crucial, because the corresponding Bragg scattering at $|\mathbf{Q}| = 0.037 \text{ Å}^{-1}$ is confined to a tiny volume of reciprocal space, making it easy to track the magnetic order as a function of pressure. The experiments were made possible through use of a single-crystal sample of exceptional structural perfection, which permits the

resolution of the magnetic satellites at high precision. Cold neutron triple-axis diffraction was performed at the Laboratoire Léon Brillouin (LLB), and small-angle neutron scattering (SANS) at the Hahn-Meitner Institut (HMI), Berlin, using a miniature clamp-type pressure cell (see Methods).

At ambient pressure, we observe resolution-limited magnetic Bragg reflections of a coherence length $\xi > 2,000$ Å at $Q = 0.037 \text{ Å}^{-1}$ (111), which are characteristic of conventional three-dimensional long-range order. This is in excellent agreement with previous studies^{19,20}. In the temperature versus pressure (T - p) plane, the magnetic ordering temperature of MnSi at ambient pressure, $T_c = 29.5$ K, determined by resistivity and susceptibility measurements^{10,11}, falls monotonically with pressure, and vanishes at $p_c = 14.6$ kbar in a first-order transition, indicative of a collapse of the static ordered magnetic moments¹⁰ (see Fig. 1). For the pressure range of interest, the lattice constant changes by only a few tenths of a per cent only in agreement with the compressibility¹⁴, and at all pressures the lattice mosaic spread remains unchanged, and is that of an essentially perfect single crystal.

Shown in Fig. 1 is a qualitative illustration of our key result, as we approach and enter the NFL phase. Below a crossover temperature T_0 and even for pressures well above p_c (that is, deep inside the NFL

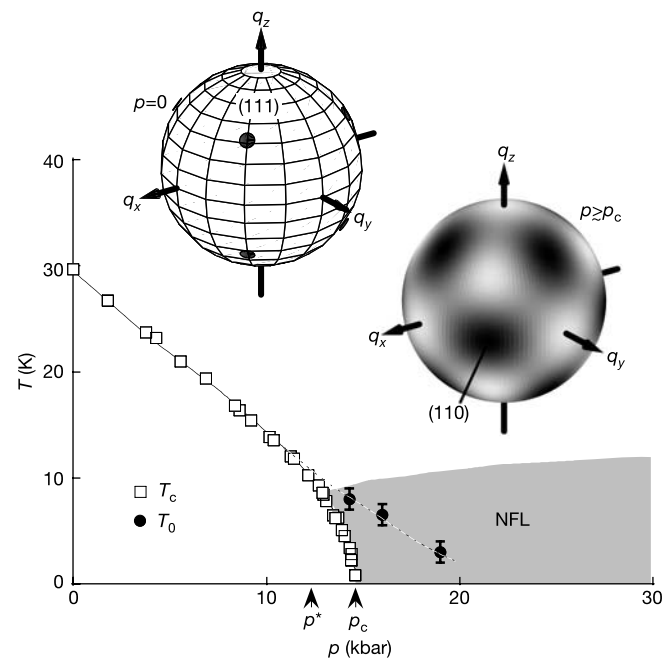


Figure 1 Schematic temperature T versus pressure p phase diagram of MnSi, and qualitative illustration of the scattering intensity characteristic of the magnetic state. Data points of $T_c(p)$ are taken from ref. 10. In the T - p plane, the transition at T_c decreases with increasing pressure and changes from second-order to weakly first-order at $p^* = 12$ kbar, before it disappears above $p_c = 14.6$ kbar. It has been argued that above p_c the non-Fermi-liquid (NFL) resistivity $\rho \propto T^{1.5}$ (which has been observed down to a few mK near p_c) provides evidence of an extended NFL phase (shaded) in a clean three-dimensional metal^{7–9}. The insets qualitatively show the location and key features of elastic magnetic scattering intensity in reciprocal space at ambient pressure (left) and at high pressure (right). Data were collected near the [110] lattice Bragg peak. At ambient pressure, resolution-limited magnetic Bragg peaks (indicated by the black dots) are observed at a distance $Q = 0.037 \text{ Å}^{-1}$, characteristic of three-dimensional long-range magnetic order, in perfect agreement with previous work. Note that the size of the dots does not reflect the resolution, and is only chosen for clarity. At high pressure, intensity is observed below a crossover temperature T_0 on the surface of a tiny sphere of radius $Q \approx 0.043 \text{ Å}^{-1}$. The intensity on this surface varies as roughly depicted by the shading, and is highest around $\langle 110 \rangle$. In the phase diagram, we show that T_0 decreases with increasing pressure, but remains finite at 19 kbar ($> p_c$), the highest pressure studied.

phase), we observe static magnetic order within the limit of our energy resolution of $50 \mu\text{eV}$. This alone is unexpected, because the bulk properties clearly suggest the loss of long-range order at p_c . Moreover, instead of well-defined magnetic Bragg satellites, scattering intensity is found essentially everywhere on the surface of a tiny sphere with radius $Q \approx 0.043 \text{ \AA}^{-1}$. The total integrated scattering intensity over the sphere at 1.6 K is of similar magnitude to that at ambient pressure. This is consistent with the observation that the bulk magnetic moment is almost independent of pressure in a polarizing magnetic field of 0.6 T (ref. 12), and suggests that we are seeing the dominant feature of the magnetic state at high pressure.

To explore the distribution of intensity on this sphere, we have performed longitudinal (radial) and transverse (tangential) scans of the magnetic intensity. For longitudinal scans in all directions, we always find resolution-limited correlation lengths $\xi > 2,000 \text{ \AA}$, in agreement with long-range longitudinal order. This is illustrated for $\langle 110 \rangle$ in Fig. 2a at $p = 14.3 \text{ kbar}$ and in Fig. 2c at $p = 16 \text{ kbar}$, where the longitudinal width does not change with temperature. In striking contrast, the corresponding transverse scans shown in Fig. 2b and d display a very large angular spread of intensity around $\langle 110 \rangle$. For 14.3 kbar, the transverse spread increases somewhat with decreasing temperature, whereas it remains almost constant for 16 kbar. The transverse width is hence not driven thermally as would be expected if there was long-range order at $\langle 110 \rangle$ with a transition temperature below our experimental range.

The detailed variation of the peak intensity with temperature for $\langle 111 \rangle$ at ambient pressure (Fig. 3a), is typical of a second-order phase transition, in excellent agreement with previous studies^{19,20}. At 1.6 K and $p = 14.3 \text{ kbar}$, only 4% of the peak intensity of ambient pressure is left (inset of Fig. 3a). Slow thermal cycles show hystereses, with $T_{\text{dn}} = 2.2 \text{ K}$ in down sweeps and $T_{\text{up}} = 4.8 \text{ K}$ in up sweeps. This is consistent with $T_c = 3.3 \text{ K}$, measured for the unchanged pressure cell in a vibrating sample magnetometer. The hysteresis is also consistent with previous reports of a first-order transition at T_c for $p^* \approx 12 \text{ kbar} < p < p_c$ (ref. 10) and itinerant metamagnetism above p_c (ref. 11) inferred from the uniform susceptibility and resistivity.

With decreasing temperature, longitudinal peak intensity appears for $\langle 110 \rangle$ below T_0 (Fig. 3b). In comparison to the sharp phase transition for $\langle 111 \rangle$, T_0 rather marks a crossover. Up to 19 kbar, the highest pressure measured, T_0 gradually falls below the onset of the NFL regime (Fig. 1), but remains finite even far above p_c .

For an interpretation of T_0 , we note that no corresponding signatures are seen in either the a.c. susceptibility or resistivity. This is surprising because, on the one hand, the magnetic intensity below T_0 is very large and, on the other hand, the electrical resistivity is exceptionally sensitive to the development of static magnetic structures, as seen by its pronounced drop at T_c (refs 7–10) and at the itinerant metamagnetic transition^{7,9}. A possible

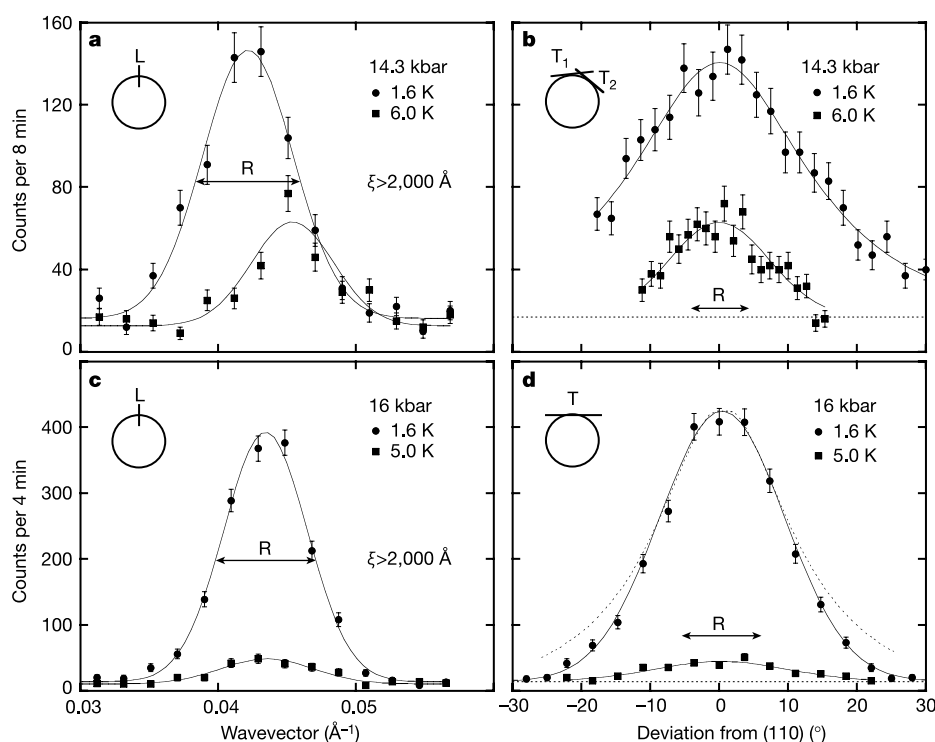


Figure 2 Longitudinal (L) and transverse (T) scans of the magnetic scattering intensity with respect to $\langle 110 \rangle$. Data are shown for 14.3 kbar (**a**, **b**), just below p_c , and 16 kbar (**c**, **d**), well above p_c . The direction of the scans with respect to a cross-section of the spherical surface is depicted in the top left corner of each panel. Note that the data at 14.3 kbar and 16 kbar are taken on two different samples with different scattering planes, thus underscoring the reproducibility of the effects (see Methods). The arrow marked R shows the instrumental resolution. **a**, Typical longitudinal scan along $\langle 110 \rangle$ at $p = 14.3 \text{ kbar}$ for 1.6 K and 6.0 K. The magnitude of $|\mathbf{Q}| \approx 0.043 \text{ \AA}^{-1}$ corresponds to a wavelength of $150 \text{ \AA}$. The resolution-limited coherence length $\xi > 2,000 \text{ \AA}$ is characteristic of long-range order. For different directions of the longitudinal scans, the intensity peaks at the same value of $|\mathbf{Q}|$, that is, the intensity lies on the surface of a sphere. **b**, Typical transverse scan at $p = 14.3 \text{ kbar}$ with respect to $\langle 110 \rangle$ in the plane

spanned by $\langle 110 \rangle$ and $\langle 111 \rangle$. The data shown is composed of two transverse scans T_1 and T_2 that were chosen such that the data displays the intensity along the arc for fixed $|\mathbf{Q}|$ with respect to $\langle 110 \rangle$. With decreasing temperature, the angular spread slightly increases. The large transverse width does not correspond to a short transverse correlation length, as discussed in the text. **c**, Typical longitudinal scan along $\langle 110 \rangle$ at $p = 16 \text{ kbar}$ for 1.6 K and 5.0 K. As for 14.3 kbar, the longitudinal width is resolution-limited. **d**, Typical transverse scan at $p = 16 \text{ kbar}$ with respect to $\langle 110 \rangle$ as above. Data were recorded along a single straight trace T. The dashed line is the variation of the intensity along a transverse arc for fixed $|\mathbf{Q}|$ (compare with Fig. 2b) calculated from the experimental data. Within the errors for the calculated curve, the transverse spread at 14.3 kbar and 16 kbar is the same. At 16 kbar, the transverse spread is roughly independent of temperature.

solution of this apparent contradiction is that the quasi-static magnetic structure seen in neutron diffraction below T_0 in fact fluctuates on the timescale relevant for transport, and that above T_0 the magnetic structures fluctuate too rapidly for large pressures and temperatures to be captured by our quasi-elastic resolution of 50 μeV . We believe that these fluctuations are the key feature of the entire NFL phase. However, studies of finely powdered MnSi may indicate that local moments exist below T_0 (ref. 24) which are static on timescales relevant for nuclear magnetic resonance (NMR) experiments. It will be a major challenge for future inelastic neutron experiments (and complementary NMR investigations) to clarify how T_0 changes with neutron energy, and whether its pressure dependence is controlled by a quantum phase transition of a subtle magnetic order at an extrapolated pressure of ~ 22 kbar, where T_0 might vanish (Fig. 1).

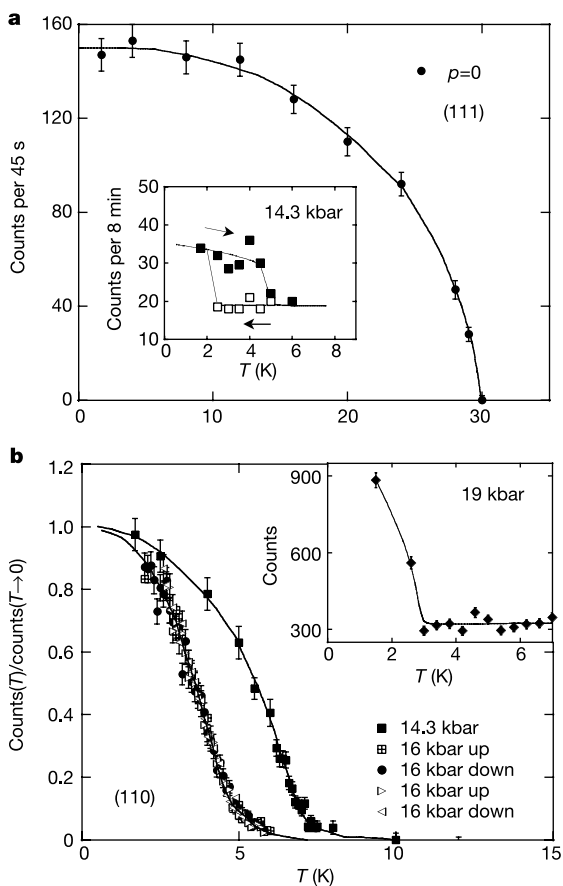


Figure 3 Temperature dependence of the longitudinal peak intensity for scans along (111) and (110) at various pressures. **a**, Temperature dependence of the satellite along (111). At ambient pressure, a second-order phase transition is observed in accordance with previous work. The inset shows the temperature dependence at 14.3 kbar, just below p_c . Taking differences of sample volume into account, only 4% of the peak intensity is left at 1.6 K. Thermal cycles display hysteresis, where T_c is in general agreement with the signature in the bulk magnetization of the same pressure cell and sample measured in a vibrating sample magnetometer. **b**, Temperature dependence for the peak intensity along (110). At 14.3 kbar, a gradual crossover below $T_0 \approx 8$ K is observed. Note that the quasi-static intensity for this satellite already appears above T_c . NFL resistivity is observed roughly between T_0 and T_c as indicated by the grey shading in Fig. 1 above T_c . At 16 kbar, the crossover temperature is shifted to $T_0 \approx 6$ K. No hysteresis is observed in thermal cycles. The inset displays the temperature dependence of the peak intensity along (110) at 19 kbar where the crossover is shifted to $T_0 \approx 3$ K. Note that the temperature-dependent data shown in the inset were taken on a small-angle spectrometer, causing a higher background. To improve statistics, data at 19 kbar were also recorded in a field-cooled state.

Before our study it was generally believed that the static amplitude of the magnetic moment (and therefore also the spiral itself) disappears at $p_c = 14.6$ kbar (refs 7–12). In contrast, the pronounced longitudinal order shows that the helix continues to exist well above p_c with a large integrated scattering intensity. For different directions of the longitudinal scans, the intensity peaks at the same value of $|Q|$; that is, the intensity lies on a sphere. Hence the transverse width must not be mistaken for short-range transverse magnetic correlations, but results from a considerable variation of the directions of the spirals. This clearly establishes that the magnetic quantum phase transition seen in the bulk properties at p_c results from an unlocking of the spiral from $\langle 111 \rangle$, driven by the weakest of the three characteristic scales discussed above.

The variations of the direction of the spirals might exist in any combination of two extremes. First, the magnetic structure may break up into a multi-domain state where the direction of the spiral varies strongly from domain to domain, and the helix may end abruptly between domains. This implies the proliferation of so-called topological defects, such as domain walls or line defects. We speculate that either the structure of those defects, the quantum fluctuations within a domain or their mutual frustration prevents the direction of the helix from locking. Second, a situation might exist where the spirals remain intact, but meander such that they do not exhibit strict directional order. Compared to the first model, such a state is defined by the absence of certain types of topological defects, and is said to develop ‘topological order’. Lammert, Rokhsar and Toner²⁵ showed that such topological order may exist in the isotropic phase of a nematic liquid crystal when the energy costs of forming disclinations are unfavourably high.

Quite generally, it is interesting to compare liquid crystals with the partial order we observe here—to our knowledge, for the first time—as a property of the conduction electrons in a metal. Owing to the lack of inversion symmetry, the properties of MnSi are related to cholesteric liquid crystals²⁶. In these cholesterics, a helical phase undergoes a sequence of phase transitions to so-called ‘blue phases’, which are characterized by complex patterns of the order parameter woven from topological defects of the underlying chiral helices²⁵. In our context, the blue phase III (also known as ‘blue fog’) is particularly interesting as it exhibits complex local order but no long-range order²⁷.

The unlocking of the spiral implies an entirely surprising microscopic origin of the NFL phase, anticipated neither experimentally nor theoretically. The identification of the corresponding low-lying excitations and their connection with the $T^{1.5}$ resistivity therefore defines a major challenge for future theoretical and experimental work. Within a partially ordered state, unconventional scattering mechanisms may, for instance, result from (pseudo-) Goldstone modes connected with fluctuations of the direction of the spirals, from internal excitations of domain walls²⁸ or other topological defects, from highly frustrated glassy magnetic configurations²⁹, from hidden quantum criticality, or from Berry phases of electrons moving on top of spin textures.

Partial electronic order as a general phenomenon of new states of matter in metals has been extensively studied theoretically in the context of the high- T_c copper oxides^{4–6} and heavy-fermion systems¹³. The clear evidence for microscopic partial order in combination with the pronounced macroscopic NFL behaviour and well-separated energy and length scales makes MnSi unique, and provides strong experimental support for the genuine existence of such phases in extremely clean systems. □

Methods

Neutron diffraction

Neutron diffraction experiments were carried out on the cold thermal neutron triple-axis spectrometer 4F1 at the LLB. A feasibility study was also carried out at the SANS spectrometer V4 at the HMI. Use of a triple-axis spectrometer allowed a significant

reduction of the diffuse scattering by the miniature pressure cell described below. On the triple-axis spectrometer, we achieved a high resolution in energy better than $50 \mu\text{eV}$. For most of the work reported here, the spectrometer was used in an elastic mode where the initial wavevector was chosen as $k_i = 1.2 \text{ \AA}^{-1}$. Data were collected near the [110] lattice Bragg point. The spectrometer 4F1 permits scans on a cartesian grid only, so that the transverse spread could only be studied in scans tangential (T) to the sphere of intensity shown in Fig. 1 (see also Fig. 2 legend). The resolution function was determined at the [110] nuclear Bragg peak, and is marked in the plots as R. For the experiments, the miniature pressure cell was mounted on a so-called orange cryostat, for minimum temperatures of 1.6 K. The data collected at the SANS spectrometer V4 reproduce the triple-axis data in the region of data overlap. We follow the usual convention where [...] denotes reciprocal lattice points, (...) a particular direction in reciprocal space and (...) the family of all equivalent directions in reciprocal space.

High-pressure techniques

High pressures were generated with a miniature clamp cell (outer diameter 12 mm) made of Cu:Be using standard PTFE sealing caps and Fluorinert as pressure transmitter. The diameter of the cylindrical samples was smaller than the inner diameter of the PTFE sealing caps by a few tenths of a millimetre. The samples were left floating free in the pressure medium to avoid the effects of differential contraction by a sample mount. Two pressure cells were prepared with a sample each, where the $\langle 110 \rangle$ and $\langle 111 \rangle$ axes were, respectively, parallel to the cylinder axis to save experiment time and to test for reproducibility. The lattice mosaic spread was found to be unchanged for both samples at all pressures studied. The pressure technique is identical to that used in a large number of studies of bulk properties of, for example, the resistivity, magnetization and quantum oscillatory phenomena, and was previously used in similar neutron diffraction studies³⁰. None of these studies indicated any hints of non-hydrostatic pressure components. However, we observe a trend for the domains in the scattering plane to be populated most strongly, that is perpendicular to the vertical axis of the pressure cell. This may be a consequence of tiny non-hydrostatic components that freeze into the pressure transmitter that we were not able to detect in the lattice mosaic or lattice form factor. For the two pressure cells, the longitudinal and transverse spread described in the text is reproducibly the same, clearly establishing that the effects are intrinsic to the sample and not caused by any tiny non-hydrostaticities. The pressure was consistently determined from (1) the change of the lattice constant, (2) the superconducting transition of Sn and (3) the bulk magnetic transition measured in a vibrating sample magnetometer when $p < p_c$.

Sample preparation and quality

The single-crystal samples investigated here were grown by r.f. induction of zone-refined high-purity starting materials on a water-cooled Cu-crucible in an ultrahigh-vacuum environment. The resulting ingots were annealed for 2 weeks near the melting point. Single-crystal samples were oriented by Laue X-ray diffraction, and cut by low-power spark erosion. The samples investigated have a tiny, resolution-limited mosaic spread $\eta \ll 0.2^\circ$, which compares with industrial quality silicon single crystals and is orders of magnitude smaller than typically found in intermetallic compounds. The very low value of η underscores the exceptional intrinsic spread around $\langle 110 \rangle$ of the magnetic state above p_c .

Received 30 September; accepted 14 November 2003; doi:10.1038/nature02232.

- Chandra, P., Coleman, P., Mydosh, J. A. & Tripathi, V. Hidden orbital order in the heavy fermion metal URu_2Si_2 . *Nature* **417**, 831–834 (2002).
- Grosche, F. M. G. *et al.* Anomalous low temperature states in CeNi_2Ge_2 and CePd_2Si_2 . *J. Phys. Condens. Matter* **12**, L533–L540 (2000).
- Dagotto, E., Hotta, T. & Moreo, A. Colossal magnetoresistant materials: The key role of phase separation. *Phys. Rep.* **344**, 1–153 (2001).
- Kivelson, S. A. *et al.* How to detect fluctuating order in the high-temperature superconductors. *Rev. Mod. Phys.* **75**, 1201–1241 (2003).
- Sachdev, S. Order and quantum phase transitions in the cuprate superconductors. *Rev. Mod. Phys.* **75**, 913–932 (2003).
- Zaanen, J. *et al.* The geometric order of stripes and Luttinger liquids. *Phil. Mag. B* **81**, 1485–1531 (2001).
- Pfleiderer, C., Julian, S. R. & Lonzarich, G. G. Non-Fermi liquid nature of the normal state of itinerant-electron ferromagnets. *Nature* **414**, 427–430 (2001).
- Pfleiderer, C. The non-Fermi liquid puzzle of MnSi at high pressure. *Physica B* **328**, 100–104 (2003).
- Doiron-Leyraud, N. *et al.* Fermi-liquid breakdown in the paramagnetic phase of a pure metal. *Nature* **425**, 595–599 (2003).
- Pfleiderer, C., McMullan, G. J., Julian, S. R. & Lonzarich, G. G. Magnetic quantum phase transition in MnSi under hydrostatic pressure. *Phys. Rev. B* **55**, 8330–8338 (1997).
- Thessieu, C. *et al.* Field dependence of the magnetic quantum phase transition in MnSi. *J. Phys. Condens. Matter* **9**, 6677–6687 (1997).
- Koyama, K., Goto, T., Kanomata, T. & Note, R. Observation of an itinerant metamagnetic transition in MnSi under high pressure. *Phys. Rev. B* **62**, 986–991 (2000).
- Barzykin, V. & Gorkov, L. P. Possibility of nontrivial magnetic order by elastic neutron scattering. *Phys. Rev. Lett.* **70**, 2479–2482 (1993).
- Fawcett, E., Maita, J. P. & Wernick, J. H. Magnetoelastic and thermal properties of MnSi. *Int. J. Magn.* **1**, 29–34 (1970).
- Bloch, D., Voiron, J., Jaccarino, V. & Wernick, J. H. The high field–high pressure magnetic properties of MnSi. *Phys. Lett. A* **51**, 259–291 (1975).
- Ishikawa, Y. *et al.* Paramagnetic spin fluctuations in the weak itinerant-electron ferromagnet MnSi. *Phys. Rev. B* **31**, 5884–5893 (1985).
- Yasuoka, H., Jaccarino, V., Sherwood, R. C. & Wernick, J. H. NMR and susceptibility studies of MnSi above T_c . *J. Phys. Soc. Jpn* **44**, 842–849 (1978).
- Taillefer, L., Lonzarich, G. G. & Strange, P. The band magnetism of MnSi. *J. Magn. Magn. Mater.* **54–57**, 957–958 (1986).

- Ishikawa, Y. & Arai, M. Magnetic phase diagram of MnSi near critical temperature studied by neutron small angle scattering. *J. Phys. Soc. Jpn* **53**, 2726–2733 (1984).
- Lebech, B. in *Recent Advances in Magnetism of Transition Metal Compounds* (eds Kotani, A. & Suzuki, N.) 167–178 (World Scientific, Singapore, 1993).
- Mena, F. P. *et al.* Heavy carriers and non-Drude optical conductivity in MnSi. *Phys. Rev. B* **67**, 241101(R) (2003).
- Bak, P. & Jensen, M. H. Theory of helical magnetic structures and phase transitions in MnSi and FeGe. *J. Phys. C* **13**, L881–L885 (1980).
- Nakanishi, O., Yanase, A., Hasegawa, A. & Kataoka, M. The origin of the helical spin density wave in MnSi. *Solid State Commun.* **35**, 995–998 (1980).
- Thessieu, C. *et al.* Pressure effect on MnSi: An NMR study. *J. Magn. Magn. Mater.* **177–181**, 609–610 (1998).
- Lammert, P. E., Rokhsar, D. S. & Toner, J. Topology and nematic ordering. *Phys. Rev. Lett.* **70**, 1650–1653 (1993).
- Wright, D. C. & Mermin, N. D. Crystalline liquids: The blue phases. *Rev. Mod. Phys.* **61**, 385–432 (1989).
- Koistinen, E. P. & Keyes, P. H. Light-scattering study of the structures of blue phase III. *Phys. Rev. Lett.* **74**, 4460–4463 (1995).
- Vonsovskij, S. V. *Magnetism* Vol. 2, Part 3 (Wiley, New York, 1974).
- Mydosh, J. *Spin Glasses: An Experimental Introduction* (Taylor and Francis, London, 1993).
- Huxley, A. *et al.* Co-existence of superconductivity and ferromagnetism in actinide compounds. *J. Phys. Condens. Matter* **15**, S1945–S1956 (2003).

Acknowledgements We acknowledge discussions with P. Böni, A. N. Bogdanov, E. Dormann, B. Fåk, P. C. Howell, B. Keimer, B. Lebech, G. G. Lonzarich, I. Mazin, A. J. Millis, K.-H. Müller, J. Mydosh, J. Kübler, B. Rössli, S. Sachdev, S. S. Saxena, J. Schmalian, Q. Si, M. Turlakov, U. Rössler, M. Vojta, P. Wölfle and J. Zaanen. Help with the experiments from M. Uhlarz, B. Hennion, J. Haug, E. García-Matres and the staff of the Laboratoire Léon Brillouin (Saclay) and the Hahn-Meitner Institut (Berlin), respectively, is also acknowledged. This work was supported by the Deutsche Forschungsgemeinschaft, the Helmholtz Gemeinschaft and the European Science Foundation under FERLIN.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.P. (Christian.Pfleiderer@physik.uni-karlsruhe.de).

Observational evidence of a change in radiative forcing due to the indirect aerosol effect

Joyce E. Penner¹, Xiquan Dong² & Yang Chen¹

¹Department of Atmospheric, Oceanic, and Space Sciences, University of Michigan, Ann Arbor, Michigan 48109-2143, USA

²Department of Atmospheric Science, University of North Dakota, Grand Forks, North Dakota 58202-9006, USA

Anthropogenic aerosols enhance cloud reflectivity by increasing the number concentration of cloud droplets, leading to a cooling effect on climate known as the indirect aerosol effect. Observational support for this effect is based mainly on evidence that aerosol number concentrations are connected with droplet concentrations, but it has been difficult to determine the impact of these indirect effects on radiative forcing^{1–3}. Here we provide observational evidence for a substantial alteration of radiative fluxes due to the indirect aerosol effect. We examine the effect of aerosols on cloud optical properties using measurements of aerosol and cloud properties at two North American sites that span polluted and clean conditions—a continental site in Oklahoma with high aerosol concentrations, and an Arctic site in Alaska with low aerosol concentrations. We determine the cloud optical depth required to fit the observed shortwave downward surface radiation. We then use a cloud parcel model to simulate the cloud optical depth from observed aerosol properties due to the indirect aerosol effect. From the good agreement between the simulated indirect aerosol effect and observed surface radiation,

we conclude that the indirect aerosol effect has a significant influence on radiative fluxes.

The Southern Great Plains (SGP) site is centred at 36° 37' N, 97° 30' W, in Oklahoma, and is a typical continental site with high aerosol number concentration (500–600 cm⁻³ for particles with diameter between 0.1 and 10 µm, and 1,000–10,000 cm⁻³ for total condensation nuclei (CN) concentration⁴) and composition of mixed-continental origin. The North Slope of Alaska (NSA) site is located at 71° 19' N, 156° 37' W, in Barrow, Alaska, and is a typical Arctic site with low aerosol concentration (CN concentration of 99 ± 120 cm⁻³ for our study) and composition determined partly by sea-salt particles (see Table 1). Here, these data and data from the Atmospheric Radiation Measurement (ARM) programme are used to examine the effects of aerosols on cloud radiative properties.

Dong *et al.*⁵ developed a method to derive cloud optical depth (τ_c) by requiring that a delta-2-stream radiative transfer model match the observed downward broadband shortwave flux at the surface, and validated its use by comparison with *in situ* aircraft data^{6–9}. Figure 1 shows the relationship between the derived τ_c and the cloud liquid water path (LWP) at both the SGP and NSA sites. In this figure, the LWP is derived from microwave radiometer brightness temperatures at 23.8 and 31.4 GHz using a statistical retrieval method¹⁰, and the cloud optical depth is derived from the radiative transfer model and Eppley precision spectral pyranometer measurements at the ARM sites^{5,11}. There is a significant difference in the slope of the relationship between cloud optical depth and LWP at these two sites. Below, we show that this difference can be explained by the indirect aerosol effect.

The time periods in this study were from 1 May to 30 September 2000 at the ARM NSA site, and from November 1996 to November 1998 at the ARM SGP site. During these two time periods the cloud temperatures at both sites are above -13 °C for most of the time, so that clouds at both sites can be viewed mostly as liquid-phase and liquid-dominant mixed-phase clouds.

We used a warm-cloud adiabatic parcel model¹² to examine the effect of aerosols on cloud drop number concentration. The parcel model requires that we specify the aerosol size distribution, total number concentration, composition, and updraft velocity. Then, we estimated the cloud optical depth using the measured cloud properties (cloud geometric thickness and LWP). Cloud geometric thickness is the difference between the cloud top height (derived from millimetre cloud radar measurements) and the cloud base height (derived from laser ceilometer measurements⁹).

We established five criteria for choosing the clouds under which cloud properties can be estimated using the retrieval method. These are: (1) only single-layer and overcast low-level stratus clouds are present as determined from cloud radar observations, (2) the cosine of solar zenith angle (μ_0) is larger than 0.2, (3) the range of solar transmission is between 0.1 and 0.7, (4) LWP is between 20 and 600 g m⁻², and (5) cloud top height is less than 3 km. We identified approximately 300 h of the stratus clouds (more than 3,600 samples at 5-min resolution) that satisfied these five criteria at the NSA site and 500 h at the SGP site. In addition to these criteria, for examining the effects of aerosols on clouds we also required that the atmosphere be well mixed below cloud base and that the cloud top height

be less than 2.5 km. These additional criteria reduced the number of cases for further analysis to 28 at the SGP site and 109 at the NSA site.

The criterion of a well-mixed atmosphere below cloud base enabled us to calculate the aerosol number concentration at cloud base by using the measured surface concentrations assuming that the aerosol mixing ratio is constant with altitude¹³. This criterion was satisfied by requiring that the water vapour mixing ratio measured by the ARM rawinsonde sounding near the cloud base and that near the surface were within ±15%. Rawinsonde data are available for each day near local noon.

The cloud droplet number concentration was calculated from the adiabatic cloud parcel model using the measured aerosol ion composition and total mass at both sites (available approximately daily at each site; ref. 14 and P. K. Quinn, personal communication) as well as the CN number concentration available from NOAA Climate Monitoring and Diagnostics Laboratory measurements¹⁵. No direct measurements of aerosol size distribution were available. However, typical continental and marine size distributions are available from a number of measurements (see Table 5.1 in ref. 1). Moreover, for the SGP site, ARM provided measurements of the total aerosol concentration with diameter >0.1 µm. So we developed a 2-mode lognormal size distribution that fitted both the total CN concentrations and the number concentration with diameter

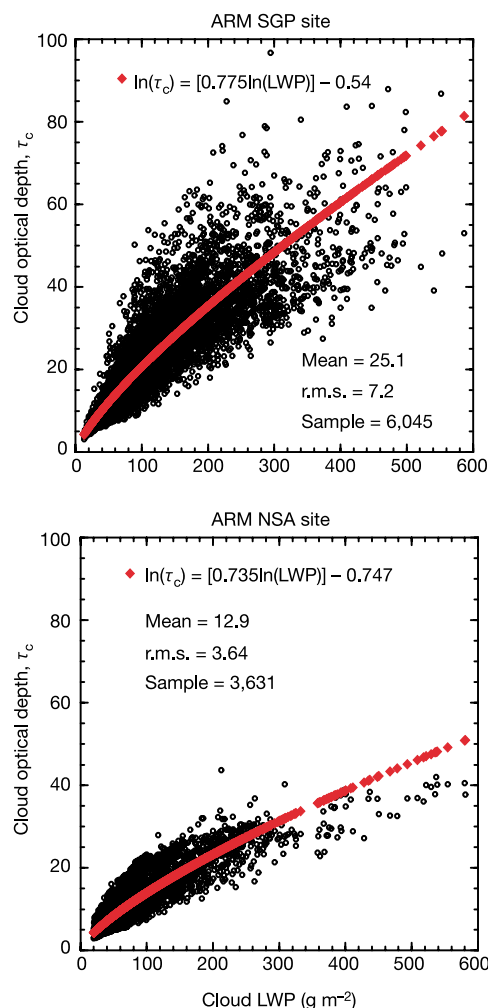


Figure 1 Cloud optical depth, as determined from the observed shortwave flux, plotted against LWP. The cloud optical depth was determined by requiring that a delta-2-stream radiative transfer model fit the observed ground level broadband shortwave radiative flux.

Table 1 Average measured aerosol composition

	SGP site		NSA site	
	µg m ⁻³	Percentage	µg m ⁻³	Percentage
NaCl	0.13	0.88	0.48	29.92
(NH ₄) ₂ SO ₄	4.36	29.10	0.54	33.28
Insoluble	9.79	65.26	0.51	31.63
NO ₃	0.40	2.69	0.03	1.88
K ⁺	0.10	0.63	0.01	0.57
Mg ²⁺	0.03	0.17	0.02	1.49
Ca ²⁺	0.19	1.25	0.02	1.21

$>0.1 \mu\text{m}$. We assumed that mode 1 had a mean diameter equal to $0.02 \mu\text{m}$ and a geometric standard deviation of 1.1, while mode 2 had a mean diameter of $0.08 \mu\text{m}$ and a geometric standard deviation of 1.75. For the NSA site, based on aerosol composition, we initially used either a typical marine size distribution¹ (a single-mode lognormal size distribution with a mean diameter equal to $0.19 \mu\text{m}$ and a geometric standard deviation of 1.5) or a typical continental size distribution¹. For these size distributions and the measured composition, a fit to the standard relationship¹⁶ for cloud condensation nuclei (CCN) as a function of supersaturation (S), $\text{CCN} = cS^k$, yields $c = 371 \pm 172$ and $k = 0.39 \pm 0.11$ for the SGP site, and $k = 0.31 \pm 0.31$ for the NSA site (c was not computed for the NSA site because almost all aerosols are activated at this site). We estimated the vertical velocity for each time period based on the average updraft velocity and the turbulent kinetic energy (TKE) from the Model Output Location Time Series data provided by the National Centers for Environmental Prediction¹⁷ using $w = \bar{w} + c\sqrt{\text{TKE}}$, where $c = 0.7$ (ref. 18). The modelled cloud optical depth is not sensitive to updraft at the NSA site, and would be very similar at the SGP site had we used the maximum millimetre cloud radar

Doppler values for the updraft velocity.

Because the liquid water content within adiabatic stratiform clouds generally increases linearly with height, given the droplet number concentration N and LWP, the cloud optical depth can be calculated from $\tau_c = 1.765(k\pi)^{1/3}N^{1/3}\rho_w^{-2/3}C_w^{-1/6}\text{LWP}^{5/6}$, where ρ_w is the liquid water density, $C_w = 2\text{LWP}/H^2$, H is the cloud geometrical thickness, and k is a parameter that describes the dispersion of the cloud drops¹⁹. The CN data used in the model simulation, as well as cloud base and cloud top, and LWP, were averaged over a 1 h time period. We used $k = [(5.0 \times 10^{-4}N) + 1.18]^{-3}$, based on a fit to observations²⁰.

Figure 2 shows the cloud optical depth and LWP relationship found here using the parcel model. The line in each curve shows the relationship found in Fig. 1 needed to match the observed surface shortwave broadband flux. The points shown by the grey dots at the NSA site have CN number concentration greater than 200 cm^{-3} and used the original single-mode size distribution. Because it is likely that higher CN number concentrations are indicative of the presence of a nucleation mode in the size distribution, we added a small lognormal mode (as used at the SGP site) to the single-mode distribution at the NSA site for these points (with the average ratio between the large and small modes that was found at the SGP site). The adjustment of size distribution to the 2-mode distribution improves the fit between the parcel model results for cloud optical depth and that required to fit the shortwave broadband flux.

The large spread in the derived cloud optical depth from surface radiation measurements shown in Fig. 1 can result from effects that are not taken into account in the parcel model: entrainment, drizzle formation, and ice formation and splintering. But could any of these processes contribute to a difference in the observed slope between the two sites? Entrainment could potentially both increase and decrease the slope of the optical depth versus LWP relationship, but if it decreases the liquid water content by half at the top of the cloud it would not have a large impact at either site. The presence of ice particles at the NSA site does not greatly affect either the retrieved LWP²¹ or retrieved optical depths⁷. Preferential drizzle at the NSA site should not greatly affect the retrieved optical depth if it is less than 25% of the total LWP⁵. The fact that warm-cloud adiabatic microphysics explains the retrieved cloud optical depth supports this assumption.

Our study shows that the cloud optical depth required to fit the surface downward shortwave flux at both the SGP and NSA sites can

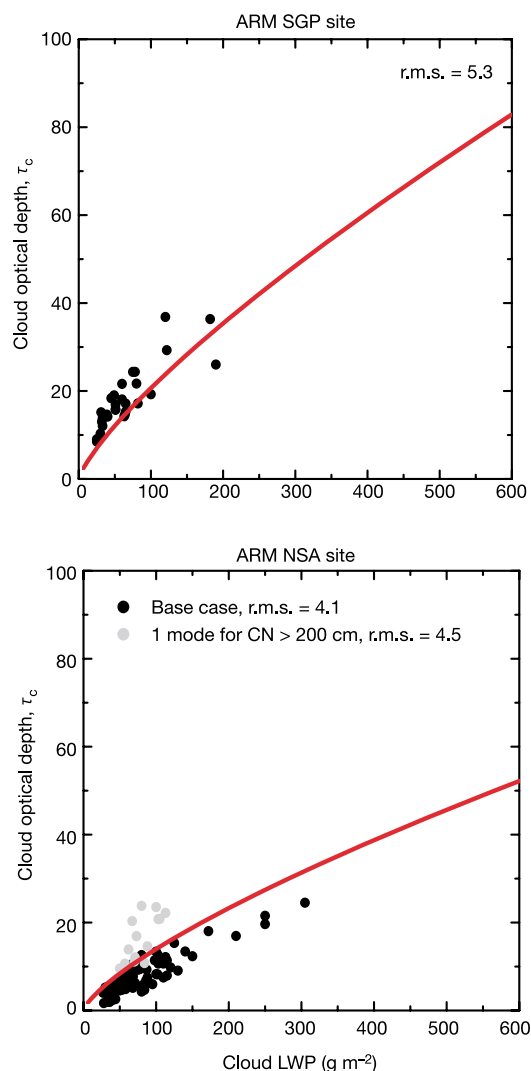


Figure 2 Cloud optical depth, as determined from the parcel model, plotted against LWP. Cloud optical depth was determined using the predicted droplet number concentrations from an adiabatic cloud parcel model and the measured aerosol concentrations and composition. In both Figs 1 and 2, the cloud LWP was that determined from the microwave radiometer measurement.

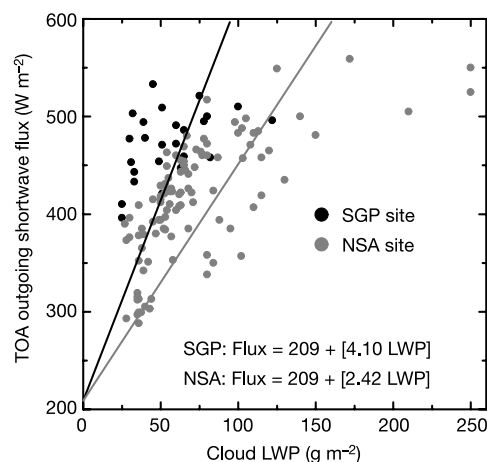


Figure 3 Modelled outgoing shortwave radiation at the SGP and NSA sites using the average albedo and zenith angle from the SGP site and the cloud properties from the adiabatic model. Only points with $\tau_c < 25$ are shown. If the anthropogenic aerosol composition at the SGP site from ref. 23 is used to determine the cloud drop concentration in the adiabatic model, then the average radiative forcing at the SGP site for these cases is 10.2 W m^{-2} . TOA, top of atmosphere.

be explained by the indirect aerosol cloud effect. The use of a parcel model to determine the cloud droplet number concentration enables us to separate the effects of the cloud LWP and cloud droplet number concentration on the cloud optical depth. An examination of the TOA shortwave flux from the radiative transfer model applied to the two sites does not directly confirm the indirect effect, because the observed surface albedos at the NSA site for our cases (0.6 ± 0.28) are significantly larger than those from the SGP site (0.2 ± 0.02). However, the model can be used to estimate the outgoing flux difference if the clouds from the NSA site had the same average surface albedo and average zenith angle as those from the SGP site (see Fig. 3). This analysis indicates that these sites provide important evidence corroborating the effect of aerosols on cloud optical properties and on shortwave fluxes at both the surface and the TOA. Moreover, the analysis indicates that a parameterization of the effects of aerosols on clouds on the basis of an adiabatic parcel model and average aerosol size distributions such as those used in current general circulation models^{18,22,23} provides a good estimate of cloud optical properties determined over a broad range of aerosol concentrations. □

Received 28 May; accepted 17 November 2003; doi:10.1038/nature02234.

1. Penner, J. E. *et al.* in *Climate Change 2001: The Scientific Basis* (eds Houghton, J. T. *et al.*) 289–348 (Cambridge Univ. Press, Cambridge, UK, 2001).
2. Rosenfeld, D. & Feingold, G. Explanation of discrepancies among satellite observations of the aerosol indirect effects. *Geophys. Res. Lett.* **30**, doi:10.1029/2003GL017684 (2003).
3. Brenguier, J.-L., Pawlowska, H. & Schüller, L. J. Cloud microphysical and radiative properties for parameterization and satellite monitoring of the indirect effect of aerosol on climate. *J. Geophys. Res.* **108**, doi:10.1029/2002JD002682 (2003).
4. Sheridan, P. J., Delene, D. J. & Ogren, J. A. Four years of continuous surface aerosol measurements from the Department of Energy's Atmospheric Radiation measurement Program Southern Great Plains Cloud and Radiation Testbed site. *J. Geophys. Res.* **106**, 20735–20747 (2001).
5. Dong, X., Ackerman, T. P., Clothiaux, E. E., Pilewskie, P. & Han, Y. Microphysical and radiative properties of stratiform clouds deduced from ground-based measurements. *J. Geophys. Res.* **102**, 23829–23843 (1997).
6. Dong, X., Ackerman, T. P. & Clothiaux, E. E. Parameterizations of microphysical and shortwave radiative properties of boundary layer stratus from ground-based measurements. *J. Geophys. Res.* **102**, 31681–31393 (1998).
7. Dong, X., Mace, G. G., Minnis, P. & Young, D. F. Arctic stratus cloud properties and their effect on the surface radiation budget: selected cases from FIRE ACE. *J. Geophys. Res.* **106**, 15297–15312 (2001).
8. Dong, X. *et al.* Comparison of stratus cloud properties deduced from surface, GOES, and aircraft data during the March 2000 ARM Cloud IOP. *J. Atmos. Sci.* **59**, 3265–3284 (2002).
9. Dong, X. & Mace, G. G. Profiles of low-level stratus cloud microphysics deduced from ground-based measurements. *J. Atmos. Ocean. Tech.* **20**, 42–53 (2003).
10. Liljegren, J. C., Clothiaux, E. E., Mace, G. G., Kato, S. & Dong, X. A new retrieval for liquid water path using a ground based microwave radiometer and measurements of cloud temperature. *J. Geophys. Res.* **106**, 14485–14500 (2001).
11. Dong, X. & Mace, G. G. Arctic stratus cloud properties and radiative forcing derived from ground-based data collected at Barrow Alaska. *J. Clim.* **16**, 445–461 (2003).
12. Liu, X. & Seidl, W. Modeling study of cloud droplet nucleation and in-cloud sulfate production during the Sanitation of the Atmosphere (SANA) 2 campaign. *J. Geophys. Res.* **103**, 16145–16158 (1998).
13. Delene, D. J. & Deshler, T. Vertical profiles of cloud condensation nuclei above Wyoming. *J. Geophys. Res.* **106**, 12579–12588 (2001).
14. Quinn, P. K. *et al.* A three-year record of simultaneously measured aerosol chemical and optical properties at Barrow, Alaska. *J. Geophys. Res.* **D 107**, doi:10.1029/2001JD001248 (2002).
15. Climate Modeling and Diagnostics Laboratory data archive (<http://www.cmdl.noaa.gov/info/ftpdata.html>) (2000).
16. Twomey, S. The nuclei of natural clouds formation. Part II: The supersaturation in natural clouds and the variation of cloud droplet concentration. *Geophys. Pura Appl.* **43**, 243–249 (1959).
17. Rogers, E., Deaven, D. G. & DiMego, G. J. The regional analysis system for the operational “early” eta model: original 80-km configuration and recent changes. *Weath. Forecast.* **10**, 810–825 (1995).
18. Lohmann, U., Feichter, J., Chuang, C. C. & Penner, J. E. Prediction of the number of cloud droplets in the ECHAM GCM. *J. Geophys. Res.* **104**, 9169–9198 (1999).
19. Brenguier, J.-L. *et al.* Radiative properties of boundary layer clouds: droplet effective radius versus number concentration. *J. Atmos. Sci.* **57**, 803–821 (2000).
20. Liu, Y. & Daum, P. H. Indirect warming effect from dispersion forcing. *Nature* **419**, 580–581 (2002).
21. Lin, B., Wielicki, B., Minnis, P. & Rossow, W. Estimation of water cloud properties from satellite microwave, infrared, and visible measurements in oceanic environments. 1. Microwave brightness temperature simulations. *J. Geophys. Res.* **103**, 3873–3886 (1998).
22. Ghan, S. J., Easter, R. C., Hudson, J. & Breon, F.-M. Evaluation of aerosol indirect radiative forcing in MIRAGE. *J. Geophys. Res.* **106**, 5317–5334 (2001).
23. Chuang, C. C. *et al.* Cloud susceptibility and the first aerosol indirect forcing: Sensitivity to black carbon and aerosol concentrations. *J. Geophys. Res.* **D 107**, doi:10.1029/2000JD000215 (2002).

Acknowledgements We thank P. Quinn for providing the composition data at the ARM SGP and NSA sites. During this study, X.D. was also supported by the NASA CERES project. This work was supported by the DOE ARM programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.E.P. (penner@umich.edu).

Tungsten isotope evidence that mantle plumes contain no contribution from the Earth's core

Anders Scherstén¹, Tim Elliott¹, Chris Hawkesworth¹ & Marc Norman²

¹Department of Earth Sciences, University of Bristol, Will's Memorial Building, Queen's Road, Bristol BS8 1RJ, UK

²Research School of Earth Sciences, The Australian National University, Canberra, Australian Capital Territory 0200, Australia

Osmium isotope ratios provide important constraints on the sources of ocean-island basalts, but two very different models have been put forward to explain such data. One model interprets ¹⁸⁷Os-enrichments in terms of a component of recycled oceanic crust within the source material^{1,2}. The other model infers that interaction of the mantle with the Earth's outer core produces the isotope anomalies and, as a result of coupled ¹⁸⁶Os–¹⁸⁷Os anomalies, put time constraints on inner-core formation^{3–5}. Like osmium, tungsten is a siderophile (‘iron-loving’) element that preferentially partitions into the Earth's core during core formation but is also ‘incompatible’ during mantle melting (it preferentially enters the melt phase), which makes it further depleted in the mantle. Tungsten should therefore be a sensitive tracer of core contributions in the source of mantle melts. Here we present high-precision tungsten isotope data from the same set of Hawaiian rocks used to establish the previously interpreted ¹⁸⁶Os–¹⁸⁷Os anomalies and on selected South African rocks, which have also been proposed to contain a core contribution⁶. None of the samples that we have analysed have a negative tungsten isotope value, as predicted from the core-contribution model. This rules out a simple core–mantle mixing scenario and suggests that the radiogenic osmium in ocean-island basalts can better be explained by the source of such basalts containing a component of recycled crust.

Many ocean-island basalts are thought to be the surface expression of mantle plumes that originate from a boundary layer within the mantle. Ocean-island basalts are geochemically distinct from mid-ocean-ridge basalts, and the differences are often attributed to recycled components of crust or ancient melt-depleted lithosphere^{7,8}. It has been argued that the Re–Os system provides powerful evidence for recycled components because Re/Os ratios are high in melts and correspondingly low in residues, which with time develop increasingly radiogenic and depleted Os-isotope compositions respectively^{1,2,9,10}. Yet Re–Os is also fractionated by inner-core crystallization, where Os partitions into the inner core leaving the outer core with a high Re/Os ratio. Thus, small core contributions are an alternative explanation for the occurrence of radiogenic Os in ocean-island basalts¹¹.

In principle, such core contributions can be tested using combined ¹⁸⁶Os–¹⁸⁷Os isotopes, because Pt–Os fractionates in the same way as Re–Os during inner-core crystallization and ¹⁹⁰Pt decays to ¹⁸⁶Os. In view of the long half-lives of ¹⁹⁰Pt ($T_{1/2} \approx 450$ Gyr) and

^{187}Re ($T_{1/2} \approx 42$ Gyr), Pt–Os and Re–Os fractionation would need to have taken place early in Earth's history to produce significant anomalies in the Os isotope system. Coupled ^{186}Os – ^{187}Os anomalies for Hawaiian picrites and Gorgonian komatiites have been attributed to small core contributions^{3–5}. If correct, this puts major constraints on the origins of such plumes and core evolution: (1) providing geochemical evidence for mantle plumes originating at the core–mantle boundary, and (2) requiring inner-core crystallization start early at 3.5 Gyr ago or earlier⁵. Independently verifying the presence of core material in the source of ocean-island basalt thus has great significance.

A new test for core interaction is now available from W isotopes and the extinct ^{182}Hf – ^{182}W system. Hf and W are both refractory elements, but because Hf is lithophile it will remain in the silicate Earth whereas the siderophile W will preferentially partition into the metallic core. Hence the silicate Earth is highly depleted in W with respect to C1 chondrite. This can be expressed in terms of the ratio of W to a refractory lithophile element of similar incompatibility (such as Th; ref. 12); C1 chondrite has $W/\text{Th} \approx 3.2$, whereas silicate Earth has W/Th_{SE} between 0.14–0.26, with a best estimate of 0.19 (ref. 12). ^{182}Hf ($T_{1/2} \approx 9$ Myr) was extant at the formation of the Solar System (t_0) and recent data show that chondritic W-isotope ratios are two epsilon units (ϵW , see Fig. 1 for definition) lower than silicate Earth^{13–15} (Table 1), which is zero by definition. This implies that Earth's core formation occurred ~ 10 – 30 Myr after t_0 depending on whether a continuous or a simple two-stage differentiation model is assumed¹³. Regardless of the timing or style of core formation, mass balance requires the core to have a $\epsilon\text{W} = -2.1$ if bulk Earth is assumed to be chondritic and silicate Earth $W/\text{Th}_{\text{SE}} = 0.19$. This value varies by only $\sim 0.1 \epsilon\text{W}$ for residual silicate Earth W/Th ratios in the range 0.14–0.26. Thus, the W-isotope budget of the Earth is well established. Critically, differences in W isotopes are imparted only at the beginning of Earth history and further differentiation processes do not affect them. This is not true for the long-lived Pt–Os and Re–Os systems, as parent–daughter fractionation at any time in Earth history will yield variable present-day Os isotopic anomalies depending on time, degree of fractionation and parent half-life. Taken together, these features make W isotopes particularly powerful as an independent test of the core-contribution model.

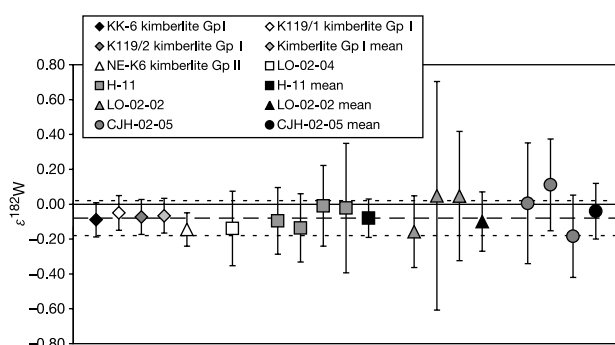


Figure 1 ϵW values for Group I and Group II kimberlites and Hawaiian picrites. ϵW was calculated using the average for the NIST SRM3163 standard solution for each analytical session: $\epsilon\text{W} = \left(\frac{^{182}\text{W}/^{184}\text{W}}{^{182}\text{W}/^{184}\text{W}} \right) \times 10,000$. To ensure that anomalies were not introduced by our chemical procedure, the NIST SRM3163 was prepared using our standard chemical separation procedure and treated as an unknown. The average $\epsilon\text{W} = -0.08 \pm 0.10$ (2σ , $n = 6$) for the chemically treated SRM3163 is plotted with dashed lines and the reproducibility is regarded as the minimum error for unknowns. In the case of the kimberlite data, the internal precision supersedes the reproducibility of the chemically treated SRM3163 and errors are therefore increased to $\pm 0.1 \epsilon\text{W}$. Although the chemically treated SRM3163 is overlapping with the untreated standard SRM3163 solution, there is a systematic shift towards lower ϵW values.

We present W-isotope ratios for Hawaiian picrites and selected South African Group I and II kimberlites (Table 1). The Hawaiian samples are the same picrites for which a core-contribution has been invoked and so our results can be compared directly with the published Os isotope data^{3,4}. Kimberlites were analysed as W isotope anomalies have been reported in these rocks and also attributed to core interactions⁶. The Hawaiian samples are tholeiitic picrites that erupted on the flanks of the Hawaiian volcanoes. In addition to Os isotopes, these samples have been analysed for a range of elements and isotope systems^{16–18}. Kimberlite samples were selected from the major collection at the University of Cape Town to cover a range of Group I and Group II samples.

All measured ϵW values are within analytical error of our measurements of NIST SRM3163, the standard widely used as the silicate Earth reference (Fig. 1). Thus we find no obvious evidence for a core contribution in either the picrites or kimberlites.

We now explore two mixing models to investigate whether Os isotopic ratios might have been influenced by core addition with no significant effect on the W isotope ratios (Fig. 2). The $^{186}\text{Os}/^{188}\text{Os}$ ratio for the outer core is taken to be 0.119870 (ref. 3), and the key controlling factors in these models are then the terrestrial inventories of W and Os and their distribution between the outer and inner core, mantle and continental crust. The bulk Earth non-volatile element abundances are assumed to be 1.85 times C1 chondrite in both models¹⁹, and it should be noted that lower estimates would straighten the mixing hyperbolae and so increase

Table 1 W isotope data for samples and NIST SRM3163 standard solution

Sample	Session	$\epsilon^{182}\text{W}$	$\pm 2\sigma$	$\epsilon^{183}\text{W}$	$\pm 2\sigma$
Kimberlites					
KK-6 kimberlite Gp I	2	-0.09	0.10	-0.06	0.08
K119/1 kimberlite Gp I	3	-0.05	0.10	-0.04	0.06
K119/2 kimberlite Gp I	3	-0.07	0.10	-0.03	0.04
NE-K6 kimberlite Gp II	3	-0.15	0.10	0.03	0.06
Hawaiian picrites					
LO-02-02	4	-0.16	0.21	-0.04	0.15
LO-02-02 repl 1	4	0.05	0.66	-0.16	0.36
LO-02-02 repl 2	4	0.05	0.37	0.06	0.26
LO-02-02 weighted average		-0.10	0.17	-0.03	0.12
LO-02-04	2	-0.14	0.21	0.07	0.17
H-11	1	-0.10	0.19	0.15	0.19
H-11 repl 1	1	-0.14	0.20	0.09	0.19
H-11 repl 2	1	-0.01	0.23	0.09	0.14
H-11 repl 3	2	-0.02	0.37	-0.08	0.27
H-11 weighted average		-0.08	0.11	0.08	0.09
C/JH-02-05	1	0.01	0.35	0.05	0.19
C/JH-02-05 repl 1	1	0.11	0.26	0.07	0.18
C/JH-02-05 repl 2	1	-0.18	0.24	-0.10	0.19
C/JH-02-05 weighted average		-0.04	0.16	0.01	0.10
Allende whole rock	5	-1.90	0.29	0.09	0.20
Replicate	5	-1.90	0.23	0.26	0.26
Coahuila IIAB iron	5	-3.49	0.10	-0.04	0.06
SRM after chemistry					
SRM 01	1	-0.20	0.18	-0.01	0.13
SRM 02	1	0.09	0.18	-0.01	0.14
SRM 03	4	-0.08	0.15	-0.09	0.09
SRM 05	4	-0.14	0.17	0.10	0.13
SRM 07	4	0.04	0.19	-0.08	0.13
SRM 08	4	-0.11	0.10	0.00	0.07
SRM weighted average		-0.08	0.10	-0.02	0.07
SRM standard solution					
		$^{182}\text{W}/^{184}\text{W}$	$\pm 2\sigma$	$^{183}\text{W}/^{184}\text{W}$	$\pm 2\sigma$
$n = 11$	1	0.864880	(3)	0.467115	(2)
$n = 13$	2	0.864812	(4)	0.467109	(4)
$n = 12$	3	0.864804	(6)	0.467122	(2)
$n = 17$	4	0.864775	(6)	0.467126	(1)
$n = 11$	5	0.864790	(7)	0.467137	(2)

ϵW data for South African Group I and Group II kimberlites and Hawaiian lavas relative to the weighted average for the untreated stock NIST SRM3163 standard solution for each analytical session. Amplifier gain calibration was made before all sessions except the first. Absolute $^{182}\text{W}/^{184}\text{W}$ and $^{183}\text{W}/^{184}\text{W}$ for NIST SRM3163 for each session are given at the bottom of the table and with 2σ errors for the last significant digit in brackets.

the expected effect on ϵW from core contribution (Fig. 2). W is siderophile during core formation and incompatible during mantle melting. Thus, the depleted mantle has low W abundances as a result of both of these processes. The W content of the core is calculated to 500 p.p.b. by mass balance using $W/Th_{SE} = 0.19$ and the reservoir masses of ref. 20. The concentration of W in the outer core, $[W]_{oc}$, is controlled by the amount of inner-core crystallization and partitioning between liquid and solid metal. We adopted the equations by Liu and Fleet²¹ and assume 2 wt% sulphur in the bulk core and 5.5 wt% inner-core crystallization yields $[W]_{oc} = 490$ p.p.b. Increasing sulphur in the core will decrease the amount of W in the outer core. In the first model, the mantle has a W/Th ratio of 0.19 and $[W]_m = 8$ p.p.b., reflecting extraction of the continental crust from the whole mantle that has subsequently been homogenized by convection²². Enrichment of W in the continental crust is estimated by comparison with Th, assuming $[Th]_{cc} = 5,600$ p.p.b. and a $W/Th_{cc} = 0.20$, which yields 1,100 p.p.b. W (refs 12, 23). Brandon *et al.*⁵ explore four different crystallization models for the inner core and we adopt the average $[Os]_{oc} = 300$ p.p.b. of their models three and four. The mantle end-member is assumed to have a $^{186}Os/^{188}Os = 0.1198345$ (ref. 3) and $[Os] = 3.1$ p.p.b. (ref. 9). The resulting mixing curve predicts $\epsilon W \approx -0.55$ for H-11, the most radiogenic Hawaiian picrite, if its elevated ^{186}Os is a result of the addition of core material. The weighted average ϵW for H-11 is clearly resolved from this mixing curve and there is therefore no support for core-derived Os using these first model parameters (Fig. 2).

As stated above, the core W isotope composition is fixed by mass balance and will not change through time. Although $^{186}Os/^{188}Os$ in the outer core and $[Os]_{oc}$ depend on the timing and extent of inner-core crystallization⁵, we have assumed that $^{186}Os/^{188}Os$ in the outer core is 0.119870, following ref. 5. If we then wish to force a core-mantle mixing model through H-11, one way is to adopt model 1 of ref. 5, where inner-core crystallization was nearly complete by 4.3 Gyr ago. Early inner-core crystallization models allow a lower

Pt/Os of the outer core and require a lower Os partition coefficient in the inner core, which implies a possible higher $[Os]_{oc} = 660$ p.p.b.: this higher $[Os]_{oc}$ will increase the mixing curvature. However, fast inner-core crystallization alone does not explain the data and a number of additional assumptions have to be made. (1) The minimum silicate Earth W-depletion to the core has to be such that $Th/W = 0.26$ and the W-extraction to the continental crust must all come from just the upper mantle, leaving the lower mantle with $[W]_{lm} = 19$ p.p.b.; (2) the core would need to contain 10 wt% sulphur, the maximum likely value; and (3) the bulk Earth composition must not have lower element abundances than our current estimate of 1.85 times C1 chondrite. Taken together these three conditions are needed to give a minimum $[W]_{oc} = 440$ p.p.b.. The resultant mixing curve (Fig. 2) is therefore tightly constrained to require fast inner-core crystallization, minimum W-depletion in silicate Earth and isolation of the lower mantle. Such an extreme model seems implausible and even so does not overlap the 2σ error of H-11.

The lack of negative ϵW anomalies in kimberlites (Fig. 1) is at odds with previous data⁶, where the most extreme anomalies are as low as some iron meteorites²⁴. If correct, this would require the core to be considerably less radiogenic than the current estimate, and to balance the silicate Earth^{13–15}, a hidden highly radiogenic silicate reservoir. Our data require no core contribution in kimberlites, nor any additional hidden reservoir, but are nevertheless compatible with growing evidence for a strong subduction component in both whole-rock kimberlites and diamond inclusions^{25,26}.

If the core-contribution model is discarded, other means of obtaining the observed Os-anomalies has to be sought. The Os-data constrains any recycled component because few crustal sources have high enough Pt/Os ratios to develop suitable ^{186}Os enrichment. Mn-nodules have $[Mn] \approx 25\%$, high $^{190}Pt/^{188}Os \approx 0.2–0.5$ (ref. 27), low $Re/Os \approx 0.1–0.5$ but high $[Os] \approx 1–3$ p.p.b. and $^{187}Os/^{188}Os \approx 0.5–1$ (ref. 28). Using these values, $\leq 2–4\%$ contamination from Mn-crusts to the peridotite source reproduces the Os isotope trends for Hawaii and Gorgona but predict no W-anomalies (Fig. 2). This model predicts a 5–9 times Mn-increase in the mantle source, and that Fe/Mn would decrease. Further detailed analyses are required to explore any correlations between Os-anomalies and Fe/Mn in the lavas, but major element compositions of erupted magmas are significantly influenced by conditions of melting, and modest variations in Mn and Fe of the source are not necessarily clearly evident in the composition of erupted melts. Brandon *et al.*⁵ argued that the intersection point for the combined ^{186}Os – ^{187}Os -trends from Gorgona, Hawaii and Siberia²⁹ may be interpreted as a single ubiquitous end-member, but we note that the range of Os compositions observed in Mn-nodules quite readily fit the uncertainty of the common intersection point reported⁵ and this constraint may therefore not be a major problem for a recycling model. Moreover a recycled origin of ^{187}Os has been implicated in Hawaiian lavas using a combined study with $\delta^{18}O$ and it was suggested that heterogeneities on the length scale of only a few kilometres are preserved in the subducted slab³⁰. If recycled Mn-crusts are confirmed to explain the ^{186}Os – ^{187}Os data, further constraints for the length scales of preservation of subducted crust are obtained because Mn-crusts make up a volumetrically small component that veneers the surface of the subducted package. □

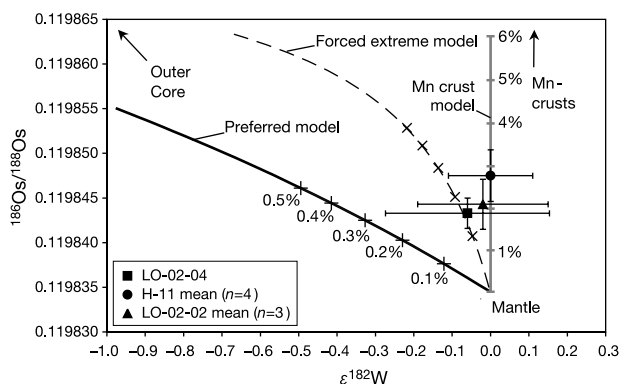


Figure 2 ϵW – $^{186}Os/^{188}Os$ mixing models for two core-contribution scenarios and for Mn-crust contamination of the Hawaiian plume source. (Os data from ref. 3.) Tick marks for the two core-contribution models represent 0.1 wt% core addition increments, whereas the tick marks for the Mn-crust model represent 1 wt% increments. Hawaiian picrites are plotted with 2σ error bars and were adjusted to the chemically treated SRM3163 by increasing their ϵW by 0.08 units to represent their relative composition more accurately. Even when errors are accounted for, the Hawaiian data are resolved from our preferred core-contribution model, although there is overlap for some of the data using more extreme parameters selected to maximize the chance of fitting. An alternative to the extreme core-contribution model is given by relatively small additions of Mn-crusts to the plume source. This model predicts no ϵW and readily fits the constraints from coupled ^{186}Os – ^{187}Os trends in Hawaiian^{1,2} and Gorgonian³ lavas. Details of the mixing parameters are discussed in the text.

Methods

Hawaiian samples were prepared from fresh interiors of newly split hand specimens and care was taken to avoid any tools containing tungsten. Kimberlite powders were prepared from macrocryst-free matrix rocks and using tungsten-free equipment at the University of Cape Town. Samples were dissolved with equal amounts of concentrated HF and HNO_3 at $120^\circ C$ for 48–72 h in Savillex vials, except for chondrites that were dissolved at $180^\circ C$ with added $HClO_4$. Vials were pre-cleaned several times in HF and HNO_3 at $200^\circ C$ to leach out W (compare with ref. 14). Samples were dried down to dryness and treated with $HClO_4$ to break down fluorides. The sample was then dissolved in HCl and diluted to 1 M with added H_2O_2 or HF to complex W before being centrifuged and loaded on a pre-cleaned

Biorad AG-1 X8 100–200 mesh column of 2–4 ml, depending on sample size. Matrix elements were removed using ten column volumes (c.v.) of the loading acid and then conditioned with 0.5 c.v. HNO_3 + 0.5 M HF. W was eluted with 7 c.v. 4 M HNO_3 + 0.5 M HF and dried to dryness. Some samples were treated with equal parts H_2O_2 and 7 M HNO_3 to break down possible organic compounds. We note that several other protocols (for example, eluting in 8 M HCl + 1 M HF) gave poor yields and in some cases, caused artificial anomalies. Samples were analysed in static mode on either of the two Finnigan Neptune plasma ionization multi-collector mass spectrometers (PIMMS) at the University of Bristol. The samples were aspirated through either a Cetac Aridus desolvating nebulizer or an Apex HF nebulizer. The data were corrected for mass bias using $^{186}\text{W}/^{183}\text{W} = 1.98594$ and screened using the corrected $^{183}\text{W}/^{184}\text{W}$ for NIST SRM3163. Total analytical blank was 200–300 pg except for the Allende chondrites where a different batch of HF gave a considerably higher blank at 940 pg. The blank contribution was insignificant to all samples (>25 ng W) except the Allende chondrite (~ 60 ng). Three samples that deviated significantly from SRM3163 in its $^{183}\text{W}/^{184}\text{W}$ were rejected. All samples are reported in Table 1 relative to stock SRM3163 using the epsilon notation for both $^{182}\text{W}/^{184}\text{W}$ and $^{183}\text{W}/^{184}\text{W}$, and where SRM3163 $\epsilon^{182}\text{W}$ and $\epsilon^{183}\text{W}$ are both zero by definition.

Received 8 July; accepted 10 November 2003; doi:10.1038/nature02221.

1. Reisberg, L. *et al.* Os isotope systematics in ocean island basalts. *Earth Planet. Sci. Lett.* **120**, 149–167 (1993).
2. Hauri, E. H. & Hart, S. R. Re-Os isotope systematics of HIMU and EMII oceanic island basalts from the south-pacific ocean. *Earth Planet. Sci. Lett.* **114**, 353–371 (1993).
3. Brandon, A. D., Walker, R. J., Morgan, J. W., Norman, M. D. & Prichard, H. M. Coupled ^{186}Os and ^{187}Os evidence for core-mantle interaction. *Science* **280**, 1570–1573 (1998).
4. Brandon, A. D., Norman, M. D., Walker, R. J. & Morgan, J. W. ^{186}Os – ^{187}Os systematics of Hawaiian picrites. *Earth Planet. Sci. Lett.* **174**, 24–42 (1999).
5. Brandon, A. D. *et al.* ^{186}Os – ^{187}Os systematics of Gorgona Island komatiites: implications for early growth of the inner core. *Earth Planet. Sci. Lett.* **206**, 411–426 (2003).
6. Collerson, K. D., Schoenberg, R. & Kamber, B. S. Unradiogenic W in kimberlites: direct evidence for core-mantle interaction. *Geochim. Cosmochim. Acta* **66**, 148 (2002).
7. Hofmann, A. W. & White, W. M. Mantle plumes from ancient oceanic crust. *Earth Planet. Sci. Lett.* **57**, 421–436 (1982).
8. McKenzie, D. & O'Nions, R. K. Mantle reservoirs and ocean island basalts. *Nature* **301**, 229–231 (1983).
9. Shirey, S. B. & Walker, R. J. The Re-Os isotope system in cosmochemistry and high temperature geochemistry. *Annu. Rev. Earth Planet. Sci.* **26**, 423–500 (1998).
10. Shafer, B. F., Turner, S., Parkinson, I., Rogers, N. & Hawkesworth, C. J. Evidence for recycled Archean oceanic mantle lithosphere in the Azores plume. *Nature* **420**, 304–307 (2002).
11. Walker, R. J., Morgan, J. W. & Horan, M. F. Osmium-187 enrichment in some plumes: evidence for core-mantle interaction? *Science* **269**, 819–822 (1995).
12. Newsom, H. E. *et al.* The depletion of tungsten in the bulk silicate Earth: constraint on core formation. *Geochim. Cosmochim. Acta* **60**, 1155–1169 (1996).
13. Yin, Q. *et al.* A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* **418**, 949–952 (2002).
14. Kleine, T., Münker, C., Mezger, K. & Palme, H. Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf-W chronometry. *Nature* **418**, 952–955 (2002).
15. Schoenberg, R., Kamber, B., Collerson, K. D. & Eugster, O. New W-isotope evidence for rapid terrestrial accretion and very early core formation. *Geochim. Cosmochim. Acta* **66**, 3151–3160 (2002).
16. Norman, M. D. & Garcia, M. O. Primitive magmas and source characteristics of the Hawaiian plume: petrology and geochemistry of shield picrites. *Earth Planet. Sci. Lett.* **168**, 27–44 (1999).
17. Bennett, V. C., Esat, T. M. & Norman, M. D. Two mantle-plume components in Hawaiian picrites inferred from correlated Os-Pb isotopes. *Nature* **381**, 221–224 (1996).
18. Bennett, V. C., Norman, M. D. & Garcia, M. O. Rhenium and platinum group element abundances correlated with mantle source components in Hawaiian picrites: sulphides in the plume. *Earth Planet. Sci. Lett.* **183**, 513–526 (2000).
19. McDonough, W. F. in *Earthquake Thermodynamics and Phase Transformations in the Earth's Interior* Vol. 76 (eds Teissieyre, R. & Majewski, E.) 3–23 (Academic, San Diego, 2001).
20. Yoder, C. F. in *Global Earth Physics: a Handbook of Physical Constants* (ed. Ahrens, T. J.) 1–31 (American Geophysical Union, Washington DC, 1995).
21. Liu, M. & Fleet, M. E. Partitioning of siderophile elements (W, Mo, As, Ag, Ge, Ga, and Sn) and Si in the Fe-S system and their fractionation in iron meteorites. *Geochim. Cosmochim. Acta* **65**, 671–682 (2001).
22. Helffrich, G. R. & Wood, B. J. The Earth's mantle. *Nature* **412**, 501–507 (2001).
23. Rudnick, R. L. & Fountain, D. M. Nature and composition of the continental crust—a lower crustal perspective. *Rev. Geophys.* **33**, 267–309 (1995).
24. Horan, M. F., Smoliar, M. I. & Walker, R. J. ^{182}W and ^{187}Re – ^{187}Os systematics of iron meteorites: Chronology for melting, differentiation, and crystallization in asteroids. *Geochim. Cosmochim. Acta* **62**, 545–554 (1998).
25. Smith, C. B. Pb, Sr and Nd isotopic evidence for sources of southern African cretaceous kimberlites. *Nature* **304**, 51–54 (1983).
26. Schulze, D. J. *et al.* Extreme crustal oxygen isotope signature preserved in coesite in diamond. *Nature* **423**, 68–70 (2003).
27. Peucker-Ehrenbrink, B. & Ravizza, G. The marine osmium isotope record. *Terra Nova* **12**, 205–219 (2000).
28. Burton, K. W. *et al.* Osmium isotope variations in the oceans recorded by Fe–Mn crusts. *Earth Planet. Sci. Lett.* **171**, 185–197 (1999).
29. Walker, R. J. *et al.* Applications of the ^{190}Pt – ^{186}Os isotope system to geochemistry and cosmochemistry. *Geochim. Cosmochim. Acta* **61**, 4799–4807 (1997).
30. Lassiter, J. C. & Hauri, E. H. Osmium-isotope variation in Hawaiian lavas: evidence for recycled oceanic lithosphere in the Hawaiian plume. *Earth Planet. Sci. Lett.* **164**, 483–493 (1998).

Acknowledgements We thank A. LeRoex, J. Gurney, K. Westerlund, N. Coe and M. Coetzee at the University of Cape Town who provided and helped us select kimberlite samples. S. Russel at the Natural History Museum, London, donated the meteorite samples. Discussions on the manuscript by E. Hauri, G. Helffrich & B. Wood are appreciated. This work is supported by a EU Marie Curie post-doctoral fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.S. (anders.schersten@bristol.ac.uk).

Fossil embryos from the Middle and Late Cambrian period of Hunan, south China

Xi-ping Dong¹, Philip C. J. Donoghue², Hong Cheng³ & Jian-bo Liu¹

¹Department of Geology, Peking University, Beijing 100871, People's Republic of China

²Department of Earth Sciences, University of Bristol, Wills Memorial Building, Queens Road, Bristol BS8 1RJ, UK

³College of Life Sciences, Peking University, Beijing 100871, People's Republic of China

Comparative embryology is integral to uncovering the pattern and process of metazoan phylogeny¹, but it relies on the assumption that life histories of living taxa are representative of their antecedents. Fossil embryos provide a crucial test of this assumption and, potentially, insight into the evolution of development, but because discoveries so far^{2–5} lack phylogenetic constraint, their significance is moot. Here we describe a collection of embryos from the Middle and Late Cambrian period (500 million years ago) of Hunan, south China, that preserves stages of development from cleavage to the pre-hatching embryo of a direct-developing animal comparable to living Scalidophora (phyla Priapulida, Kinorhyncha, Loricifera). The latest-stage embryos show affinity to the Lower Cambrian embryo *Markuelia*³, whose life-history strategy contrasts both with the primitive condition inferred for metazoan phyla and with many proposed hypotheses of affinity^{3,6}, all of which prescribe indirect development. Phylogenetic tests based on these embryological data suggest a stem Scalidophora affinity. These discoveries corroborate, rather than contradict, the predictions of comparative embryology, providing direct historical support for the view that the life-history strategies of living taxa are representative of their stem lineages.

All of the embryos are preserved in calcium phosphate with varying degrees of fidelity, from extremes in which structures of less than 0.3 μm are preserved, to others in which it is possible to make out only the general outline of the embryos; most specimens show a microspheritic surface texture indicative of bacterially mediated soft-tissue replacement⁷.

The embryos vary in size and in the developmental stages represented. The smallest (diameter 236 μm) and earliest (Fig. 1a, b) stage is a cleavage embryo preserving the surface boundaries between blastomeres. Although not complete, dividing the surface area of the embryos by the average cell area suggests 485 cells. The remaining collection of embryos (for example, Figs. 1c–f) range in diameter from 370 to 411 μm . Three further specimens (for example, Fig. 1c) are later developmental stages but do not fully occupy the volume of the embryo, and much of the surface area preserves undifferentiated tissue, probably yolk.

The latest-stage embryos have a vermiform anatomy coiled into a sphere with head and tail juxtaposed laterally (Fig. 1f). The trunk is coiled in an inverted S-shaped double loop and is annulated throughout (Fig. 1d); individual annulae are 20–25 μm in length and 180–190 μm in width, suggesting that there are ~ 130 annulae in total. The annulated surface is ornamented by perpendicularly orientated, anastomosing ribs of 0.3–0.5 μm . Annular margins are seen to extend internally in fractured specimens (Fig. 1d) and are, thus, not simply surface undulations. The head and tail are distinguished as terminal, spine-bearing regions in which annulae are less obviously developed. The tail bears six curved spines of roughly equal length (Fig. 1e, f), varying between specimens from 50 to 95 μm in length. These are broadly paired but seem to have been arranged radially about a terminal depression or opening (Fig. 1e). Four spines are positioned marginal to, and oriented away from, this depression, and the remaining two may occur in the depression itself.

The presumed head is well preserved in only one specimen (Figs 1f and 2a–f), which reveals a terminal mouth of 46 μm in diameter that is surrounded by at least three rows of partially overlapping rows of flat, broadly triangular, posteriorly directed spines of up to 27 μm in

length and 11 μm in width; the spine rows extend around all visible sides of mouth and are, thus, presumably radial (part of the subterminal region is obscured by collapse and diagenetic cement). Some spines show evidence of collapse and folding (Fig. 2c, d), suggesting that they are not solid epicuticular thickenings. Both the posterior and the oral spines show a smooth surface, distinct from the finely ribbed trunk, head and tail, indicating that they are not simple extensions of the body wall and may have been more extensively cuticularized. From anterior to posterior, the reconstructed embryo (Fig. 3a) would have been just over 3 mm in length.

The new material is closely comparable to *Markuelia secunda*³, whose embryos are also annulated, vermiform and coiled in a double 'S' loop, with head and tail juxtaposed. It differs by lacking conical trunk protuberances and by showing a broadly radial, rather than bilateral, arrangement of terminal (posterior) spines. But given the reported distinctiveness of *Markuelia* from the embryos of any well-known living taxon³, the similarities shared with the new material must reflect close kinship; indeed, we deem it sufficiently

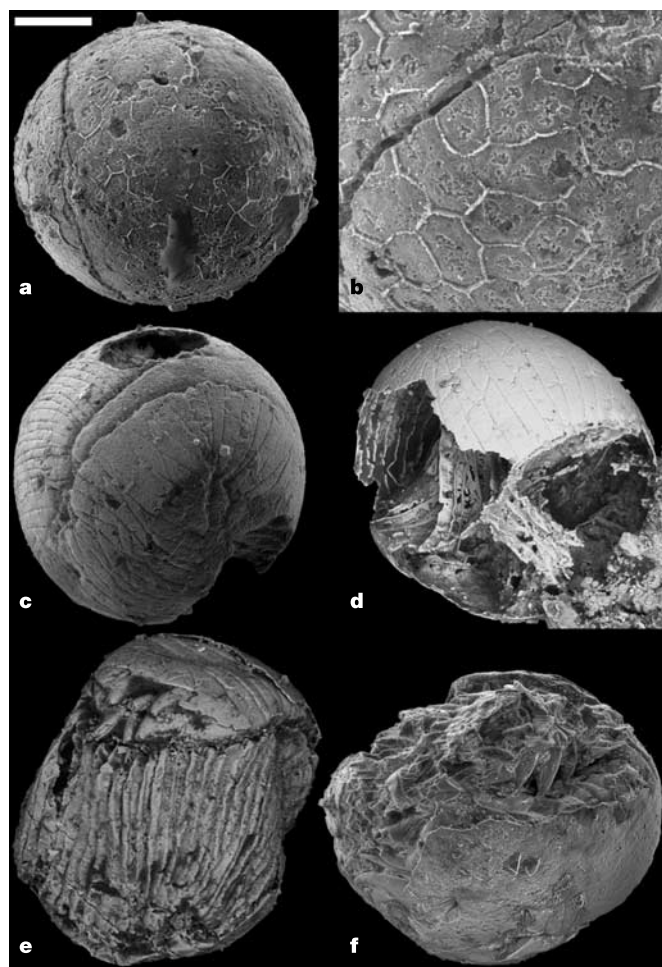


Figure 1 *Markuelia hunanensis*. **a**, **b**, Late cleavage embryo with surface margins of blastomeres preserved (GMPKU2007). **c**, Annulated embryo with yolk tissue (GMPKU2008). **d**, Fractured late-stage embryo with internal annular divisions (GMPKU2009). **e**, Late-stage embryo with tail and six terminal hooked spines (GMPKU2011). **f**, Late-stage embryo with head (upper) and tail (lower) juxtaposed (GMPKU2010). Relative scale bar, 58 μm (**a**); 104 μm (**b**); 92 μm (**c**); 27 μm (**d**); 80 μm (**e**); 88 μm (**f**).

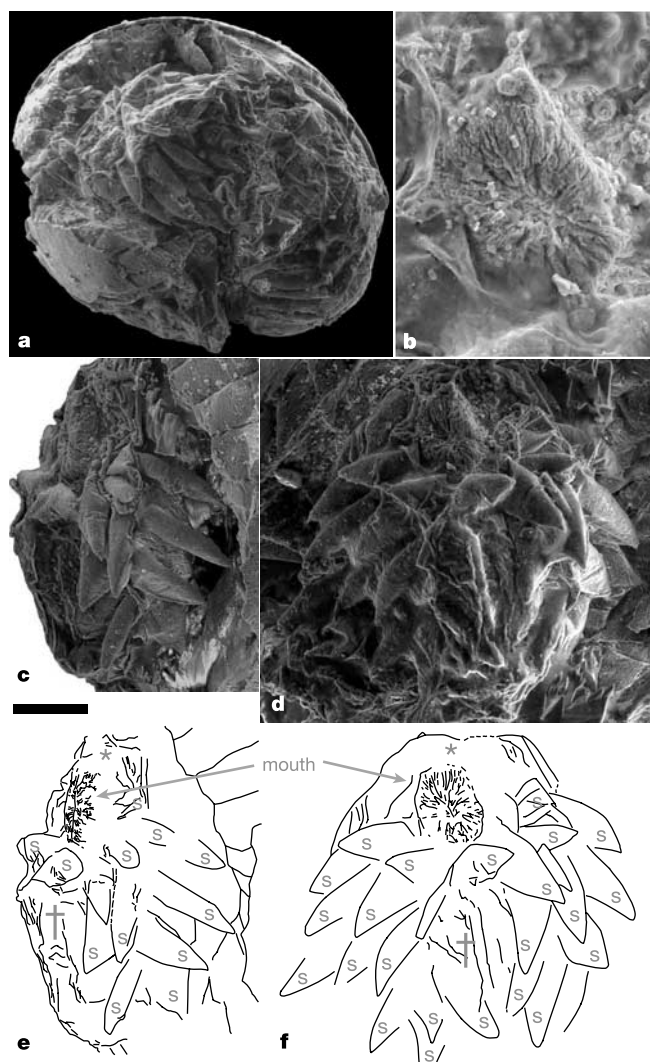


Figure 2 The oral region of the holotype specimen (GMPKU2010). **a**, Whole embryo viewed directly onto the oral region. **b**–**f**, Detail of the mouth cone (**b**), viewed from the side (**c**, **e**) and from above (**d**, **f**). The image in **a** is rotated 90° counter-clockwise from all other images. The position of scalids is shown by s; asterisks indicate the subterminal region of the mouth, which is obscured by diagenetic cement; daggers indicate the area obscured by surface coating. Relative scale bar, 34 μm (**a**); 40 μm (**b**); 11 μm (**c**).

similar to place this new species, *Markuelia humanensis*, in the same genus. Thus, the additional anatomical information provided by *M. humanensis* can be used to constrain not only its own phylogenetic relationships, but also those of the genus *Markuelia*.

The features common to *Markuelia secunda* and *M. humanensis* have hitherto been taken to suggest an affinity with lobopods³, annelids³ and/or halkieriids⁶ (a grade of organisms putatively including ancestors and close relatives of the annelids, molluscs and brachiopods⁸); the additional anatomical characters from *M. humanensis* facilitate a test of these hypotheses. The absence of paired appendages, or their anlagen, in these later-stage embryos precludes a close affinity to lobopods and annelids. An affinity with halkieriids was originally proposed on the basis of co-occurrence in geological samples, and the possibility that the trunk spines of *M. secunda* may represent scalid anlagen. The presence of oral spines demands comparison to the putative radulae of halkieriids and *Wiwaxia*. However, the adoral rather than aboral orientation of these spines, their apparent circumoral rather than bilateral disposition and, most significantly, the terminal rather than sub-terminal position of the mouth all contrast with the situation in halkieriids⁸ and *Wiwaxia*⁹. Finally, although we would not wish to preclude the possibility of identifying evolution of life-history strategies, the direct mode of development shown by *Markuelia* contrasts significantly with molluscs, brachiopods and most annelids. Thus, the affinities of *Markuelia* lie elsewhere.

The head and mouth of *Markuelia* are potentially more instructive phylogenetically: the terminal position of the mouth is a key feature of cycloneuralians *sensu*¹, and the radial arrangement of

circum-oral spines is directly comparable, both morphologically and topologically, to the scalids of Scalidophora. Circum-oral scalids are present in other cycloneuralians, such as nematodes and nematomorphs¹⁰, although these may be convergent¹¹. Nevertheless, the five phyla share an eversible proboscis (introvert) associated with scalids and are united as Introverta on the basis of this apomorphy¹. The structure of the mouth in *Markuelia*, although not definitive, is compatible with the presence of a proboscis. The other features shown by *Markuelia*, including annulation, spine-bearing trunk and terminal appendages, are also found among all five introvert phyla.

Thus, the possession of multiple rows of circum-oral scalids suggests a closer affinity of *Markuelia* to Scalidophora than to nematodes, nematomorphs or a combination thereof, or to any more inclusive groupings, although this is contingent on a phylogenetic test of character distribution. To this end, we undertook a phylogenetic analysis of Introverta, which was based on several data sets compiled to resolve the interrelationships of fossil and living introverts^{12–14}. This analysis (see Supplementary Information) resolves *Markuelia* as a sister taxon to extant Scalidophora, although in an unresolved position relative to two other stem scalidophorans; with successive weighting, *Markuelia* is resolved as the most derived member of the scalidophoran stem lineage (Fig. 3b and Supplementary Information).

Resolution of the affinity of *Markuelia* provides a basis for determining its significance in understanding the evolution of development and the inferential power of evolutionary embryology. Initial speculation suggested a lophotrochozoan affinity and, thus, evidence that the preponderance of indirect developers undergoing a planktotrophic larval stage may be an artifact of convergence⁶. The placement of *Markuelia* among Ecdysozoa, a clade of animals lacking a planktotrophic larval stage, corroborates rather than contradicts the predictions of evolutionary embryology. Nevertheless, the phylogenetic position of *Markuelia* supports the prediction that stem scalidophorans were direct developers and that the secondary larvae of priapulids and loriciferans reflect a derived, rather than plesiomorphic life-history strategy for Scalidophora and Introverta. This conclusion has previously been contingent on distant outgroups (Arthropoda and/or Gastrotricha) because of conflict in the ingroup, but the placement of *Markuelia* resolves this conflict with respect to both Scalidophora and Introverta without reference to an outgroup.

Suggestions that the similarities between nematoids and scalidophorans are convergent¹¹ are weakened by the extension of many potential homologies into the scalidophoran stem group. These include the possession of an eversible proboscis and the possession of circum-oral scalids. That the absence of giant fibres in the body wall of extant scalidophorans, a putative synapomorphy of nematoids^{11,14}, might be a secondary adaptation to small size seems to be borne out by their presence in palaeoscoleceids (P.C.J.D., unpublished observation), resolved herein among the priapulid stem lineage. Differences in the organizational symmetry of oral structures between nematoids and scalidophorans may well be resolved by further examination of the stem groups of, and to, these superphyletic clades. Introvert sympleisomorphies potentially have an even wider distribution, extending to the groundplan of Introverta plus Panarthropoda, given that a subset of these characters are general also to gastrotrichs, a sister taxon to this clade¹³. Significantly, this corroborates the view that a terminally positioned mouth is an arthropod sympleisomorphy¹⁵.

The identification of *Markuelia* as a stem scalidophoran provides both direct evidence for the establishment of this lineage in the Lower Tommotian (Lower Cambrian; 538 million years ago¹⁶) and indirect evidence for the establishment of the nematoid lineage by this time, and fulfils the prediction of a ghost lineage inferred from the earliest fossil record of the introvert sister clade Euarthropoda^{17,18}. The reconstruction of a Cambrian stem group to a clade

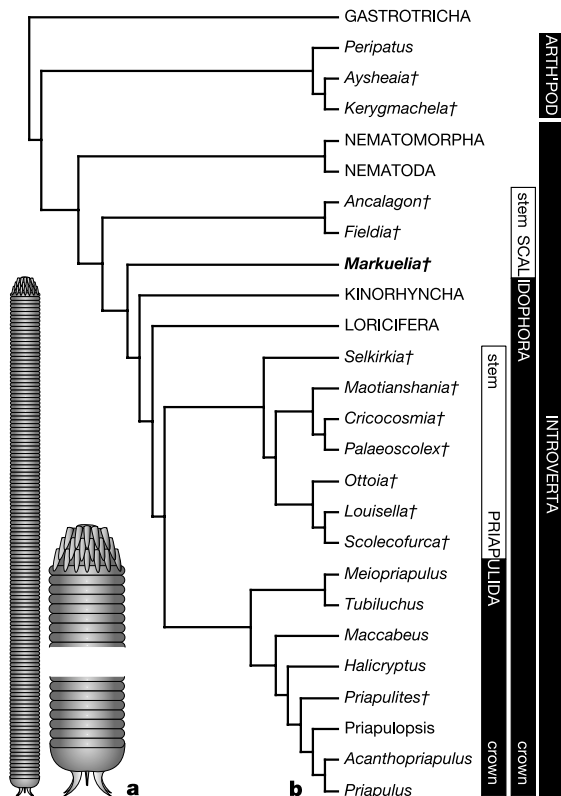


Figure 3 *Markuelia* is a stem scalidophoran. **a**, Reconstruction of the embryo unfurled. **b**, Tree summarizing interrelationships among living and fossil introverts and the outgroup, derived from numerical cladistic analysis of 87 morphological characters using branch-and-bound and heuristic search techniques of the unweighted and reweighted datasets in PAUP* 4.0b10. See Supplementary Information for component analyses and their statistical support. Daggert indicate extinct taxa.

of three phyla is compatible with the view that the establishment of phyla continued into the Cambrian¹⁹, and it provides a more complete framework of character evolution in the assembly of their bodyplans for which developmental explanations must now be sought. The placement of fossil-based developmental data in this context facilitates a thorough integration of palaeontology into evolutionary developmental biology in understanding the evolution of development.

Superphylum Introverta Nielsen 1995 (ref. 20)

Genus *Markuelia* Val'kov 1983 (ref. 21)

Markuelia hunanensis Dong & Donoghue sp. nov.

Etymology. Named for its provenance in the Chinese Province of Hunan.

Holotype. Geological Museum of Peking University, Beijing, China: GMPKU2010.

Stratigraphy and locality. Middle Upper Cambrian Bitiao Formation in Wangcun, Hunan, south China.

Diagnosis. A species of *Markuelia* with six terminal, posterior spines arranged radially and away from a central depression or opening, lacking trunk spines, and showing at least three overlapping rows of posteriorly directed circum-oral scalds.

Remarks. *Markuelia* cannot be allocated to existing rank taxa below Introverta without rendering such taxa paraphyletic. The establishment of many hierarchies of new rank taxa, solely to encompass *Markuelia*, will not serve scientific communication. □

Received 4 September; accepted 4 November 2003; doi:10.1038/nature02215.

- Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla* 2nd edn (Oxford Univ. Press, Oxford, 2001).
- Zhang, X.-G. & Pratt, B. R. Middle Cambrian arthropod embryos with blastomeres. *Science* **266**, 637–639 (1994).
- Bengtson, S. & Yue, Z. Fossilized metazoan embryos from the earliest Cambrian. *Science* **277**, 1645–1648 (1997).
- Kouchinsky, A., Bengtson, S. & Gershwil, L. Cnidarian-like embryos associated with the first shelly fossils in Siberia. *Geology* **27**, 609–612 (1999).
- Xiao, S., Zhang, Y. & Knoll, A. H. Three-dimensional preservation of algae and animal embryos in a Neoproterozoic phosphate. *Nature* **391**, 553–558 (1998).
- Conway Morris, S. Eggs and embryos from the Cambrian. *Bioessays* **20**, 676–682 (1998).
- Briggs, D. E. G., Kear, A. J., Martill, D. M. & Wilby, P. R. Phosphatization of soft-tissue in experiments and fossils. *J. Geol. Soc.* **150**, 1035–1038 (1993).
- Conway Morris, S. & Peel, J. S. Articulated halkieriids from the Lower Cambrian of north Greenland and their role in early protostome evolution. *Phil. Trans. R. Soc. Lond. B* **347**, 305–358 (1995).
- Conway Morris, S. The Middle Cambrian metazoan *Wiwaxia corrugata* (Matthew) from the Burgess Shale and *Ogygopsis* Shale, British Columbia, Canada. *Phil. Trans. R. Soc. Lond. B* **307**, 507–586 (1985).
- Bohall, P. J., Wells, M. R. & Chandler, C. M. External morphology of larvae of *Chordodes morgani* (Nematomorpha). *Invertebr. Biol.* **116**, 26–29 (1997).
- Schmidt-Rhaesa, A. Phylogenetic relationships of the nematomorpha—a discussion of current hypotheses. *Zool. Anz.* **236**, 203–216 (1998).
- Wills, M. A. Cambrian and recent disparity: the picture from priapulids. *Paleobiology* **24**, 177–199 (1998).
- Schmidt-Rhaesa, A., Bartolomaeus, T., Lemburg, C., Ehlers, U. & Garey, J. R. The position of the Arthropoda in the phylogenetic system. *J. Morphol.* **238**, 263–285 (1998).
- Lemburg, C. *Ultrastrukturelle Untersuchungen an den Larven von Halicyptus spinulosus und Priapulid caudatus—Hypothesen zur Phylogenie der Priapulida und deren Bedeutung für die Evolution der Nemathelminthes* (Cuvillier, Göttingen, 1999).
- Eriksson, B. J. & Budd, G. E. Onychophoran cephalic nerves and their bearing on our understanding of head segmentation and stem-group evolution of Arthropoda. *Arthropod Struct. Dev.* **29**, 197–209 (2000).
- Jenkins, R. J. F., Cooper, J. A. & Compston, W. Age and biostratigraphy of Early Cambrian tuffs from SE Australia and southern China. *J. Geol. Soc.* **159**, 645–658 (2002).
- Budd, G. E. & Jensen, S. in *Telling the Evolutionary Time: Molecular Clocks and the Fossil Record* (eds Donoghue, P. C. J. & Smith, M. P.) (CRC, London, 2003).
- Crimes, T. P. Trace fossils and correlation of late Precambrian and early Cambrian strata. *Geol. Mag.* **124**, 97–119 (1987).
- Budd, G. E. & Jensen, S. A critical reappraisal of the fossil record of bilaterian phyla. *Biol. Rev.* **74**, 253–295 (2000).
- Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla*, 1st edn (Oxford Univ. Press, New York, 1995).
- Val'kov, A. K. in *Pozdnij Dokembrij i Rannij Paleozoj Sibiri, Vendskie Otlozheniya* (eds Khomentovsky, V. V., Yakshin, M. S. & Karlova, G. A.) 37–48 (Inst. Geol. Geofiz. SO AN SSSR, Novosibirsk, 1983).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank W. Guo for field assistance, and S. Bengtson and S. Conway Morris for discussion. This work was supported by grants from the National Natural Science Foundation of China, Laboratory of Paleobiology and Stratigraphy of the Nanjing Institute of Geology & Palaeontology, Chinese Academy of Sciences and the Ministry of Education of China (to X.-P.D.); the Royal Society and Natural Environment Research Council (to P.C.J.D.); and the Ministry of Science and Technology of China (to J.-B.L.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to X.-P.D. (dongxp@pku.edu.cn) or P.C.J.D. (phil.donoghue@bristol.ac.uk).

Whole-lake carbon-13 additions reveal terrestrial support of aquatic food webs

Michael L. Pace¹, Jonathan J. Cole¹, Stephen R. Carpenter², James F. Kitchell², James R. Hodgson³, Matthew C. Van de Bogert¹, Darren L. Bade², Emma S. Kritzberg⁴ & David Bastviken⁵

¹Institute of Ecosystem Studies, Millbrook, New York 12545, USA

²Center for Limnology, University of Wisconsin, Madison, Wisconsin 53706, USA

³Department of Biology, St. Norbert College, De Pere, Wisconsin 54115, USA

⁴Department of Ecology/Limnology, Lund University, S-223 62 Lund, Sweden

⁵Department of Water and Environmental Studies, Linköping University, SE 581 83 Linköping, Sweden

Ecosystems are supported by organic carbon from two distinct sources. Endogenous carbon is produced by photosynthesis within an ecosystem by autotrophic organisms. Exogenous carbon is produced elsewhere and transported into ecosystems. Consumers may use exogenous carbon with consequent influences on population dynamics, predator–prey relationships and ecosystem processes¹. For example, exogenous inputs provide resources that may enhance consumer abundance beyond levels supported by within-system primary production². Exogenous fluxes of organic carbon to ecosystems are often large, but this material is recalcitrant and difficult to assimilate, in contrast to endogenously produced organic matter, which is used more easily^{3,4}. Here we show, by the experimental manipulation of dissolved inorganic ¹³C in two lakes, that internal primary production is insufficient to support the food webs of these ecosystems. Additions of NaH¹³CO₃ enriched the ¹³C content of dissolved inorganic carbon, particulate organic carbon, zooplankton and fish. Dynamics of ¹³C indicate that 40–55% of particulate organic carbon and 22–50% of zooplankton carbon are derived from terrestrial sources, showing that there is significant subsidy of these ecosystems by organic carbon produced outside their boundaries.

In lakes, grazing and microbial–detrital trophic pathways support higher consumers^{5,6}. The importance of dissolved and particulate organic matter not derived from primary producers has been recognized^{4,7}, but not widely considered in studies of food webs¹. System respiration exceeds gross primary production in many lakes, implying the metabolism of exogenous organic carbon⁸. Aquatic bacteria use exogenous dissolved organic carbon (DOC), and subsequent consumption of bacteria by predators provides a pathway for transfer into food webs^{9,10}. However, efficiencies of bacterial growth and trophic transfer of exogenous carbon are often low, and the importance of this pathway is uncertain^{6,11}. Direct use of exogenous carbon by animal consumers is also possible^{12,13}, but poorly understood.

Large-scale tests of the importance of exogenous carbon to food

webs are few; however, the elimination of forest litter inputs in small streams was found to reduce the biomass and production of consumers¹⁴. In larger systems, stable isotopes of elements such as carbon, nitrogen and sulphur can help to distinguish organic matter sources¹⁵, but often isotopic differences between endogenous and exogenous materials are not distinct. An alternative approach is to label endogenous primary production with a tracer and to investigate whether fluxes are sufficient to support consumers¹⁶.

We made daily additions of $\text{NaH}^{13}\text{CO}_3$ to the mixed layers of Paul Lake and Peter Lake over 42 d, enriching ^{13}C in DIC (Fig. 1). Phytoplankton rapidly fixed carbon from enriched DI^{13}C , as indicated by increases in the $\delta^{13}\text{C}$ of POC ($\delta^{13}\text{C}_{\text{POC}}$; Fig. 1). POC is a mixture of phytoplankton, small heterotrophic organisms (mainly bacteria) and non-living organic matter. When the $\text{NaH}^{13}\text{CO}_3$ addition ended, $\delta^{13}\text{C}$ of DIC ($\delta^{13}\text{C}_{\text{DIC}}$) and $\delta^{13}\text{C}_{\text{POC}}$ decreased towards background levels. The $\delta^{13}\text{C}$ dynamics of *Daphnia* spp., a consumer of phytoplankton and important prey of invertebrates and fish¹⁷, was very similar to $\delta^{13}\text{C}_{\text{POC}}$ in both lakes (Fig. 1).

To assess the relative contributions of endogenous and exogenous carbon to POC, we compared three models of the observed ^{13}C

dynamics. The simplest model assumes that POC is derived solely from within-lake primary production. $\delta^{13}\text{C}_{\text{POC}}$ on day t is predictable from $\delta^{13}\text{CO}_{2(\text{aq})}$ and photosynthetic fractionation (ϵ_p).

$$\delta^{13}\text{C}_{\text{POC}t} = (\delta^{13}\text{CO}_{2(\text{aq})} - \epsilon_p)_t \quad (1)$$

$\delta^{13}\text{CO}_{2(\text{aq})}$ is the ^{13}C content of aqueous CO_2 , the form of DI^{13}C that is taken up most readily by phytoplankton¹⁸. Chemical fractionation of ^{13}C between $\text{CO}_{2(\text{aq})}$ and HCO_3^- was calculated from DIC, DI^{13}C , pH and temperature¹⁹. The unknown parameter is ϵ_p , the biological fractionation of ^{13}C during photosynthesis. Fractionation is negative in relation to $\delta^{13}\text{CO}_{2(\text{aq})}$ typically ranges from 21 to 28‰ (ref. 20).

The second model assumes that POC is a mixture of phytoplankton carbon derived from photosynthesis and carbon with a terrestrial isotopic signature.

$$\delta^{13}\text{C}_{\text{POC}t} = (1 - w)(\delta^{13}\text{CO}_{2(\text{aq})} - \epsilon_p)_t + w(-28) \quad (2)$$

Terrestrial plants dominant in the lake watersheds use C_3 photosynthesis and have a $\delta^{13}\text{C}$ value close to -28 ‰. The unknown parameters in equation (2) are ϵ_p and w , the proportion of POC of terrestrial origin.

The third model was similar to model 2 but included an additional term for carbon derived from recent photosynthesis, denoted by the parameter m .

$$\delta^{13}\text{C}_{\text{POC}t} = (1 - w)[(1 - m)(\delta^{13}\text{CO}_{2(\text{aq})} - \epsilon_p)_t + m(\delta^{13}\text{CO}_{2(\text{aq})} - \epsilon_p)_{t-u}] + w(-28) \quad (3)$$

Carbon turnover is not instantaneous, and this model accounts for carbon from primary production residing over time in POC. The unknown parameters in equation (3) are ϵ_p , w , m and u , where m is the proportion of POC formed by photosynthesis u days before t (see Methods and Supplementary Information).

Assuming that POC is completely phytoplankton carbon (model 1) provides a poor fit to the data (Table 1). This model predicts more negative values of $\delta^{13}\text{C}_{\text{POC}}$ in Paul Lake than were observed at the beginning and end of the experiment, and more positive values than were observed during the middle of the experiment (Fig. 2a). Adding a second pool of carbon (that is, terrestrial carbon) not labelled by the ^{13}C addition substantially improves the model fit (Fig. 2b, Table 1). Model 3 provides the best fit to the data (Fig. 2c), and this improvement is highly significant (Table 1). Similar results were obtained when models 1, 2 and 3 were applied to data from Peter Lake (Table 1).

The best estimates of ϵ_p (± 1 s.d.) are 11.4 ± 0.7 ‰ and 11.5 ± 1.1 ‰ in Paul Lake and Peter Lake, respectively. For all

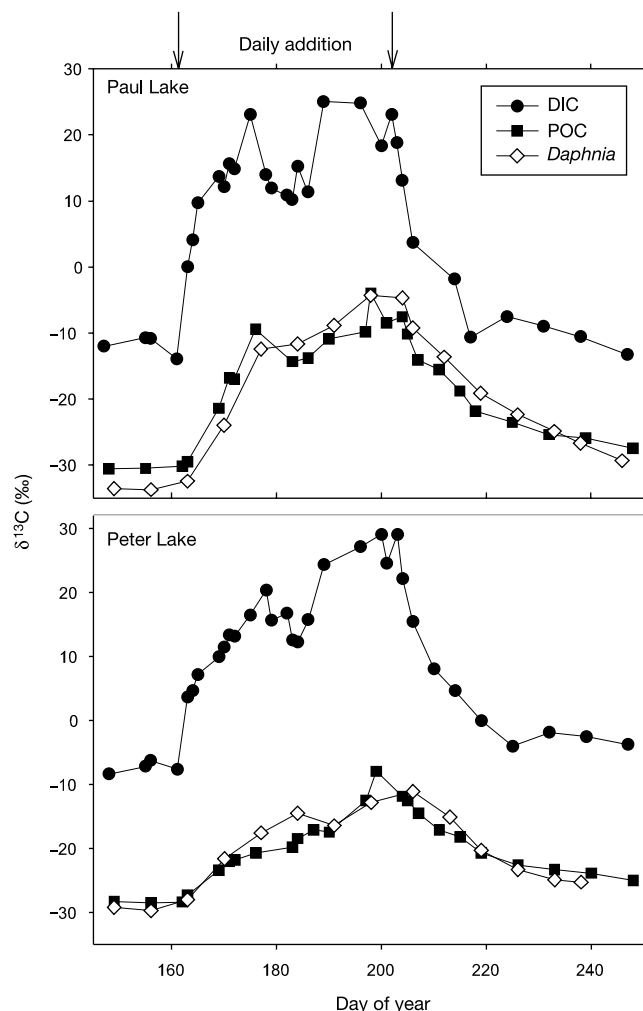


Figure 1 Dynamics of ^{13}C before, during, and after the addition of $\text{NaH}^{13}\text{CO}_3$ to Paul Lake and Peter Lake. Additions were made daily from day 162 to day 203 of the year. The carbon pools are dissolved inorganic carbon (DIC), particulate organic carbon (POC) and *Daphnia* spp. The ^{13}C content of samples is expressed as $\delta^{13}\text{C} = 1,000 \times [(R/0.011237) - 1]$, where R is the ratio of ^{13}C to ^{12}C in the sample and 0.011237 is the ratio in a standard.

Table 1 Parameter values* for models of POC and *Daphnia* ^{13}C dynamics in Paul Lake and Peter Lake

POC	Model	ϵ_p	w	m	u	RSD	AIC
Paul Lake	1	17.8 ± 1.42				7.07	83.0
Paul Lake	2	9.5 ± 2.00	0.46 ± 0.047			3.21	66.1
Paul Lake	3	11.5 ± 0.90	0.40 ± 0.027	0.44 ± 0.071	8 ± 0.87	1.58	53.0
Peter Lake	1	21.6 ± 1.55				7.53	84.5
Peter Lake	2	10.2 ± 2.44	0.59 ± 0.042			2.49	59.9
Peter Lake	3	11.4 ± 1.25	0.55 ± 0.028	0.51 ± 0.116	9 ± 1.75	1.49	51.7
<i>Daphnia</i>							
Paul Lake	1	16.6 ± 1.94				7.45	56.8
Paul Lake	2	11.3 ± 3.21	0.38 ± 0.096			5.26	53.3
Paul Lake	3	14.5 ± 0.95	0.22 ± 0.045	0.56 ± 0.077	13 ± 1.07	2.22	43.5
Peter Lake	1	19.7 ± 1.89				7.03	49.2
Peter Lake	2	10.8 ± 3.06	0.55 ± 0.067			2.86	38.6
Peter Lake	3	12.5 ± 0.72	0.50 ± 0.020	0.92 ± 0.076	6 ± 0.48	0.83	25.2
Peter Lake	3	12.8 ± 1.08	0.48 ± 0.031	0.54 ± 0.073	10 ± 1.33	1.33	30.7

*Shown are the parameter values (ϵ_p , w , m and u) with bootstrapped estimates of s.d., mean residual s.d. (RSD) and Akaike information criteria (AIC) for models of POC and *Daphnia* ^{13}C dynamics in Paul Lake and Peter Lake. Number of observations: $n = 24$, POC models; $n = 16$, Paul *Daphnia* models; $n = 14$, Peter *Daphnia* models. $\dagger u$ was fixed in this analysis.

three models, ε_p is less than 20‰ (Table 1), which contrasts with the fractionation observed in cultures and marine systems^{18,20} but is consistent with a previous whole-lake ^{13}C addition¹⁶.

We considered three analogous models for the ^{13}C dynamics of *Daphnia* spp. (hereafter *Daphnia*) by replacing the left-hand term of equations (1) to (3) with $\delta^{13}\text{C}_{\text{Daphnia}}$. Model 3 also provided the best fit to the $\delta^{13}\text{C}_{\text{Daphnia}}$ data (Table 1). For Paul Lake, the model captures the overall dynamics but predicts higher values before the addition and overestimates the decline in ^{13}C after the addition (Fig. 3, top). Measured losses of ^{13}C from both *Daphnia* and POC (Fig. 2) were slower than the model estimates, probably owing to a low turnover of structural carbon.

For Peter Lake, two lags (u) fit the data (Table 1). A 6-d lag provides the best fit (Fig. 3, bottom) but a high estimate of m . This model implies that nearly all of the phytoplankton carbon consumed by *Daphnia* is recent as opposed to produced on the same day as measurements of *Daphnia* ^{13}C . A time lag of 10 d results in a slightly worse fit but an estimate of m that is consistent with other models. In either case, model 3 provides an excellent fit (predicted versus observed: $u = 6$, coefficient of correlation squared (r^2) = 0.98; $u = 10$, $r^2 = 0.95$). The strong relationship between $\delta^{13}\text{C}_{\text{POC}}$ and $\delta^{13}\text{C}_{\text{Daphnia}}$ (Fig. 1) also suggests that *Daphnia* fed primarily in the surface layer and not on sources from deeper layers

including methanotrophic bacteria (refs 21, 22, and see Supplementary Information).

Carbon assimilated by *Daphnia* is transferred up the food web. *Daphnia* is a key prey for insects that are, in turn, principal dietary items of fish¹⁸. In addition, *Daphnia* accounts for nearly 100% of the initial diet of young-of-the-year (YOY) largemouth bass (*Micropterus salmoides*) in Paul Lake²³. During the ^{13}C addition, $\delta^{13}\text{C}$ of YOY bass increased from -28.4 to -9.7 , indicating that exogenous and endogenous carbon moves from POC to *Daphnia* to fish.

On the basis of the models, POC consisted of 40% and 55% organic matter with a $\delta^{13}\text{C}$ value close to -28 ‰ in Paul Lake and Peter Lake, respectively (model 3; 'w' in Table 1). The balance derived from primary production. POC turns over within 1–10 d through respiration, conversion to DOC, consumption or sinking¹⁶. Thus, new inputs must replace the continuous losses of POC. The ^{13}C additions indicate that a large fraction of this carbon comes from sources other than within-lake primary production.

There are several possible pathways for exogenous carbon to enter lakes and subsequently be transformed into POC available to consumers such as *Daphnia*. DOC of terrestrial origin can be converted to POC through chemical and biological mechanisms including aggregation, coagulation, flocculation and bacterial uptake. But bacterial production seems insufficient to account for much of the carbon required by consumers¹⁶. On the basis of atmospheric inputs of terrestrial material to lakes²⁴, we estimate that POC deposition could be $50\text{--}100\text{ mg C m}^{-2}\text{ d}^{-1}$, an input sufficient to account for the PO^{13}C dynamics observed in the addition experiments. Physical conversion of DOC to POC could constitute a large flux, because DOC is over ten times the concentration of POC. DOC in river water rapidly aggregates into submicrometer particles that contain DNA, lipids and sugars¹³.

Similar aggregations in lakes might constitute food for organisms such as *Daphnia* that feed on small particles. Resuspended sediments might also contribute to POC with a $\delta^{13}\text{C}$ value similar to that of terrestrial carbon. Paul Lake and Peter Lake are small and

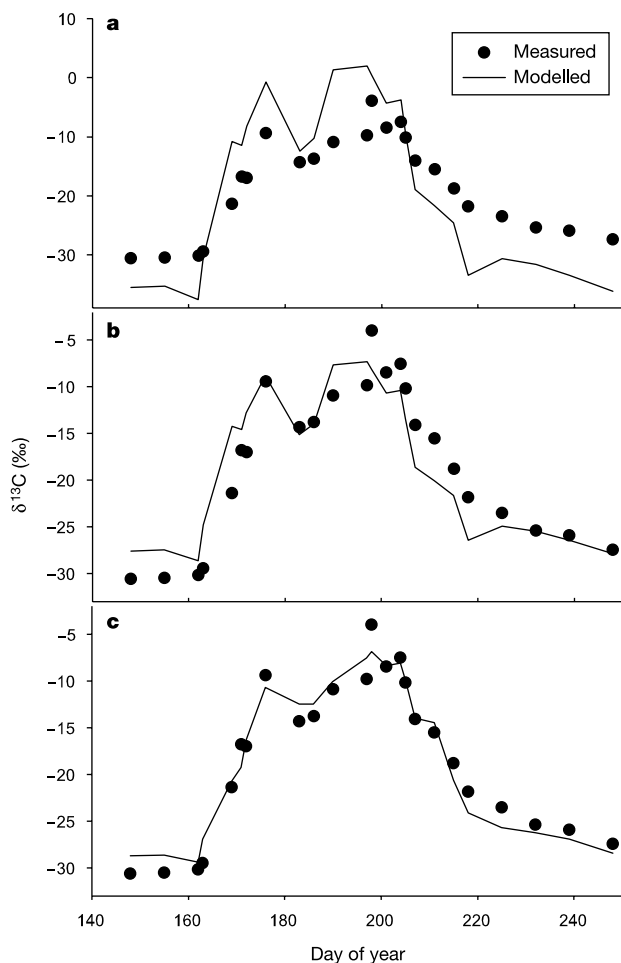


Figure 2 Comparison between measured and modelled $\delta^{13}\text{C}_{\text{POC}}$ in Paul Lake. **a**, Model 1 assumes that all POC is phytoplankton. **b**, Model 2 assumes that POC is a mixture of phytoplankton and organic carbon with a fixed value of -28 , the signature of terrestrial organic matter. **c**, Model 3 is similar to model 2 but contains additional parameters for carbon that is recently derived from primary production in the lake. Equations for each model are given in the text.

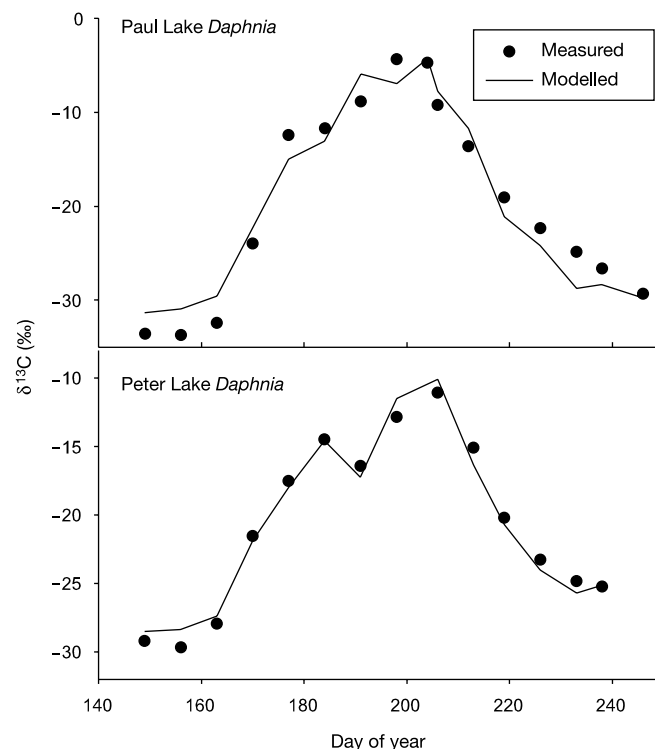


Figure 3 Comparison between measured and modelled $\delta^{13}\text{C}$ of *Daphnia* in Paul Lake and Peter Lake based on model 3.

deep, with a limited littoral area as a source of resuspended material, and limited fetch to create wind-driven waves¹⁷. Thus, resuspension is probably a small input in these lakes. A combination of terrestrial carbon inputs from both dissolved and particulate materials probably accounts for the large pool of exogenous carbon in POC. Quantitative measurements of these inputs are needed to determine their significance.

The dynamics of the ¹³C isotope in our whole-lake addition experiments clearly indicates that about half of the POC, first, is not based on current or recent primary production; second, has a $\delta^{13}\text{C}$ value consistent with that of terrestrial plants; third, is assimilated by animals; and last, is passed up the food web to fish. These results provide evidence that inputs of terrestrial carbon support secondary production in lakes of 17,000–24,000 m². The magnitude of this support is large but may be sensitive to nutrient enrichment. In a previous experiment with additions of nitrogen and phosphorus, less than 10% of zooplankton biomass seemed to be derived from allochthonous carbon⁷. Under the ambient nutrient loading of our experiments, *Daphnia*, which is traditionally considered a herbivore, received 20–50% of its carbon from a source other than primary production. Thus, the food webs are not simply based on internal primary production but are also coupled to watershed inputs of organic carbon. These inputs supplement the production of consumers and seem to be important in many ecosystems¹. For lakes, movements of organic carbon connect terrestrial ecosystems with aquatic food webs and support not only bacteria, but also invertebrates and fish. □

Methods

Study sites and field methods

Paul Lake and Peter Lake are located at University of Notre Dame Environmental Research Center, Michigan, USA (46° 13' N, 89° 13' W). These lakes have been studied extensively, and published data¹⁷ on the mixing depth, temperature, pH, carbon concentrations, productivity and trophic structure aided our design of the ¹³C additions. A third lake, Hummingbird, was monitored for ¹³C as a reference for natural variability (Supplementary Information). We dissolved 20 g (Paul) or 30 g (Peter) of NaH¹³CO₃ in lake water and pumped the solution into the upper, mixed layer from a moving boat. Previous studies²⁵ indicate that solute additions mix completely within 24 h. Loadings were 0.24 mol of ¹³C d⁻¹ to Paul Lake and 0.35 mol of ¹³C d⁻¹ to Peter Lake. The cumulative addition over 42 d altered total DIC by less than 1%. The added NaH¹³CO₃ (Isotech) was more than 99% ¹³C. Daily additions maintained an increase in DI¹³C over longer periods of time. In a previous experiment, a single addition of NaH¹³CO₃ to a lake resulted in a rapid loss of ¹³C to the atmosphere and limited movement of ¹³C into the food web, despite the addition of nutrients (N and P) to promote carbon fixation¹⁶.

Measurement of standing stocks and ¹³C

Carbon concentrations and ¹³C contents in various pools were measured at weekly or more frequent intervals. Concentrations of DIC were measured with a gas chromatograph. POC and chlorophyll *a* were measured in samples retained on GF/F filters. POC was determined in concert with analyses of ¹³C (see below). Filters for chlorophyll *a* were extracted in methanol, and concentrations were determined by fluorometry. We estimated zooplankton abundance from vertical net hauls. The ¹³C content of DIC, POC, the cladoceran, *Daphnia* spp. and YOY largemouth bass, *M. salmoides* (Paul Lake only), was measured by mass spectrometry.

We collected samples for DIC and POC daily before each addition. Samples for analysis of DI¹³C were collected in 100-ml serum vials, acidified to pH 2 with H₂SO₄, closed with gas-tight seals, and sent to the University of Waterloo stable isotope facility for analysis. Analysis of POC samples for both total C concentration and ¹³C content were conducted at the University of Alaska stable isotope facility. A subset of the daily collection of DIC and POC samples was analysed to characterize the dynamics of ¹³C before, during and after the addition. Analyses of ¹³C in dried tissue of *Daphnia* spp. and *M. salmoides* were also made at the University of Alaska. *Daphnia* were collected at least weekly by towing a net immediately beneath the surface at night and subsequently isolating animals from live samples under a dissecting scope. YOY bass were collected either from the guts of adult bass or by electroshock sampling. We dried and ground whole fish for tissue analysis.

Statistics and modelling

To derive a time series, $\delta^{13}\text{C}_{\text{CO}_2(\text{aq})}$ data were interpolated to generate daily values, and the model was evaluated for the dates that $\delta^{13}\text{C}_{\text{POC}}$ or $\delta^{13}\text{C}_{\text{Daphnia}}$ were sampled in each lake. The best-fitting parameters for the three models (Table 1) were determined by least squares²⁶, and models were compared by the mean residual s.d. and the Akaike information criterion²⁷. A worked example of model 3 calculations and the method for estimating the parameter *u* are given in the Supplementary Information. We estimated parameter uncertainty by bootstrapping²⁸. Bootstrapped parameter distributions for the models were roughly normally distributed, and parameter bias²⁸ was only a few per cent of

the s.d. Residuals for the models were approximately normally distributed. We chose a value of -28‰ to represent terrestrial organic matter on the basis of measurements of the $\delta^{13}\text{C}$ of wetland and forest plants (*n* = 10) surrounding Paul Lake and Peter Lake.

For Peter Lake and Paul Lake, model 3 residuals averaged about 1.5‰ (Table 1) and were comparable to error in the data. PO¹³C samples taken over 24 h on 20–21 June 2001 had a mean and s.d. (*n* = 6) of -16.6 ± 2.1‰ and -21.0 ± 0.7‰ for Paul Lake and Peter Lake, respectively. The variation of these samples reflects analytical and sampling error, as well as dynamics in PO¹³C (uptake and loss of ¹³C). Although the diel sampling series probably overestimates variability of the PO¹³C data used in the model (all samples were taken at the same time of day), the comparison indicates that the model residual error was of the same magnitude as the error in the data.

Received 8 September; accepted 18 November 2003; doi:10.1038/nature02227.

- Polis, G. A., Anderson, W. B. & Holt, R. D. Toward an integration of landscape and food web ecology: the dynamics of spatially subsidized food webs. *Annu. Rev. Ecol. Syst.* **28**, 289–316 (1997).
- Nakano, S. & Murakami, M. Reciprocal subsidies: dynamic interdependence between terrestrial and aquatic food webs. *Proc. Natl Acad. Sci.* **98**, 166–170 (2001).
- Cyr, H. & Pace, M. L. Magnitude and patterns of herbivory in aquatic and terrestrial ecosystems. *Nature* **361**, 148–150 (1993).
- Wetzel, R. G. Death, detritus, and energy flow in aquatic ecosystems. *Freshw. Biol.* **33**, 83–89 (1995).
- Lindeman, R. L. The trophic–dynamic aspect of ecology. *Ecology* **23**, 399–418 (1942).
- Kalff, J. *Limnology: Inland Water Ecosystems* 366–368 (Prentice-Hall, Upper Saddle River, NJ, 2002).
- Birge, E. A. & Juday, C. The organic content of the water of small lakes. *Proc. Am. Phil. Soc.* **66**, 357–372 (1927).
- Cole, J. J., Pace, M. L., Carpenter, S. R. & Kitchell, J. F. Persistence of net heterotrophy in lakes during nutrient addition and food web manipulations. *Limnol. Oceanogr.* **45**, 1718–1730 (2000).
- Hessen, D. O. in *Aquatic Humic Substances* (eds Hessen, D. O. & Tranvik, L. J.) 285–315 (Springer, New York, 1998).
- Grey, J., Jones, R. I. & Sleep, D. Seasonal changes in the importance of the source of organic matter to the diet of zooplankton in Loch Ness, as indicated by stable isotope analysis. *Limnol. Oceanogr.* **46**, 505–513 (2001).
- Tranvik, L. J. in *Aquatic Humic Substances* (eds Hessen, D. O. & Tranvik, L. J.) 259–283 (Springer, New York, 1998).
- Roditi, H. A., Fisher, N. S. & Sanudo-Wilhelmy, S. A. Uptake of dissolved organic carbon and trace elements by zebra mussels. *Nature* **407**, 78–80 (2000).
- Kerner, M. *et al.* Self-organization of dissolved organic matter to micelle-like microparticles in river water. *Nature* **422**, 150–154 (2003).
- Wallace, J. B., Eggert, S. L., Meyer, J. L. & Webster, J. R. Multiple trophic levels of a forest stream linked to terrestrial litter inputs. *Science* **277**, 102–104 (1997).
- Peterson, B. J., Howarth, R. W. & Garritt, R. H. Multiple stable isotopes used to trace organic matter in estuarine food webs. *Science* **227**, 1361–1363 (1985).
- Cole, J. J., Carpenter, S. R., Kitchell, J. F. & Pace, M. L. Pathways of organic carbon utilization in small lakes: results from a whole-lake ¹³C addition and coupled model. *Limnol. Oceanogr.* **47**, 1664–1675 (2002).
- Carpenter, S. R. & Kitchell, J. F. (eds) *The Trophic Cascade in Lakes* (Cambridge Univ. Press, Cambridge, 1993).
- Laws, E. A., Popp, B. N., Bidigare, R. R., Kennicutt, M. C. & Macko, S. A. Dependence of phytoplankton carbon isotopic composition on growth rate and [CO₂]_{aq}: theoretical considerations and experimental results. *Geochim. Cosmochim. Acta* **59**, 1131–1138 (1995).
- Mook, W. G., Bimmerson, J. C. & Staverman, W. H. Carbon isotope fractionation between dissolved bicarbonate and gaseous carbon dioxide. *Earth Planet. Sci. Lett.* **22**, 169–176 (1974).
- Bidigare, R. R. *et al.* Consistent fractionation of ¹³C in nature and in the laboratory: growth rate effects in some haptophyte algae. *Global Biogeochem. Cycles* **11**, 279–292 (1997).
- Jones, R. I., Grey, J., Sleep, D. & Arvola, L. Stable isotope analysis of zooplankton carbon nutrition in humic lakes. *Oikos* **86**, 97–104 (1999).
- Bastviken, D., Ejlertsson, J., Sundh, I. & Tranvik, L. Methane as a source of carbon and energy for lake pelagic food webs. *Ecology* **84**, 1000–1012 (2003).
- Post, D. M. *et al.* Seasonal effects of variable recruitment of a dominant piscivore on pelagic food web structure. *Limnol. Oceanogr.* **42**, 722–729 (1997).
- Cole, J. J., Caraco, N. F. & Likens, G. E. Short-range atmospheric transport: A significant source of phosphorus to an oligotrophic lake. *Limnol. Oceanogr.* **35**, 1230–1237 (1990).
- Cole, J. J. & Pace, M. L. Hydrologic variability of small, northern lakes measured by the addition of tracers. *Ecosystems* **1**, 310–320 (1998).
- Hilborn, R. & Mangel, M. *The Ecological Detective: Confronting Models with Data* 114–117 (Princeton Univ. Press, Princeton, NJ, 1997).
- Burham, K. P. & Anderson, D. R. *Model Selection and Inference: a Practical Information Theoretic Approach* (Springer, New York, 1998).
- Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgments We thank D. Fischer, J. Hinkle, C. M. Gille, C. L. Fankhauser, B. Charipar and M. Bastviken for assistance; and N. Haubenstock and R. Drimmie for help and advice on stable isotope analysis. The University of Notre Dame Environmental Research Center provided facilities and support for the field research. This research was supported by a collaborative grant from the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.P. (pacem@ecostudies.org).

Bayesian integration in sensorimotor learning

Konrad P. Körding & Daniel M. Wolpert

Sobell Department of Motor Neuroscience, Institute of Neurology,
University College London, Queen Square, London WC1N 3BG, UK

When we learn a new motor skill, such as playing an approaching tennis ball, both our sensors and the task possess variability. Our sensors provide imperfect information about the ball's velocity, so we can only estimate it. Combining information from multiple modalities can reduce the error in this estimate^{1–4}. On a longer time scale, not all velocities are a priori equally probable, and over the course of a match there will be a probability distribution of velocities. According to bayesian theory^{5,6}, an optimal estimate results from combining information about the distribution of velocities—the prior—with evidence from sensory feedback. As uncertainty increases, when playing in fog or at dusk, the system should increasingly rely on prior knowledge. To use a bayesian strategy, the brain would need to represent the prior distribution and the level of uncertainty in the sensory feedback. Here we control the statistical variations of a new sensorimotor task and manipulate the uncertainty of the sensory feedback. We show that subjects internally represent both the statistical distribution of the task and their sensory uncertainty, combining them in a manner consistent with a performance-optimizing bayesian process^{4,5}. The central nervous system therefore employs probabilistic models during sensorimotor learning.

Subjects reached to a visual target with their right index finger in a virtual-reality set-up that allowed us to displace the visual feedback of their finger laterally relative to its actual location (Fig. 1a; see Methods for details). On each movement, the lateral shift was randomly drawn from a prior distribution that was gaussian with a mean shift of 1 cm to the right and a standard deviation of 0.5 cm (Fig. 1b). We refer to this distribution as the true prior. During the movement, visual feedback of the finger position was only provided briefly, midway through the movement. We manipulated the reliability of this visual feedback on each trial. This feedback was either provided clearly (σ_0 condition, in which the uncertainty comes from intrinsic processes only), blurred to increase the uncertainty by a medium (σ_M) or large (σ_L) amount, or was withheld altogether leading to infinite uncertainty (σ_∞). Visual information about the position of the finger at the end of the movement was provided only on clear feedback trials (σ_0) and subjects were instructed to get as close to the target as possible on all trials.

Subjects were trained for 1,000 trials on the task, to ensure that they experienced many samples of the lateral shift drawn from the underlying gaussian distribution. After this period, when feedback was withheld (σ_∞), subjects pointed 0.97 $\pm$ 0.06 cm (mean $\pm$ s.e.m. across subjects) to the left of the target showing that they had learned the average shift of 1 cm experienced over the ensemble of trials (Fig. 1a, example finger and cursor paths shown in green). Subsequently, we examined the relationship between imposed lateral shift and the final location that subjects pointed to. On trials in which feedback was provided, there was compensation during the second half of the movement (Fig. 1a, example finger and cursor paths for a trial with lateral shift of 2 cm shown in blue). The visual feedback midway through the movement provides information about the current lateral shift. However, we expect some uncertainty in the visual estimate of this lateral shift. For example, if the lateral shift is 2 cm, the distribution of sensed shifts over a large number of trials would be expected to have a gaussian distribution centred on 2 cm with a standard deviation that increases with the blur (Fig. 1c).

There are several possible computational models that subjects could use to determine the compensation needed to reach the target on the basis of the sensed location of the finger midway through the movement. First (model 1), subjects could compensate fully for the

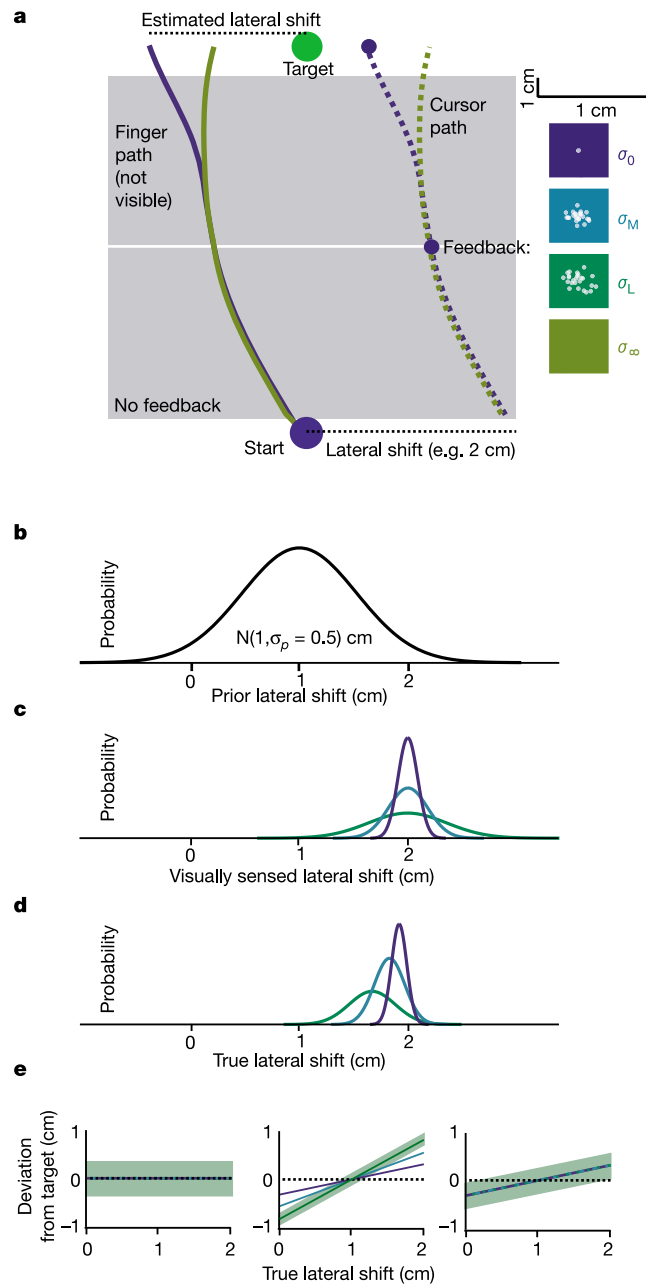


Figure 1 The experiment and models. **a**, As the finger moves from the starting circle, the cursor is extinguished and shifted laterally from the true finger location. The hand is never visible. Halfway to the target, feedback is briefly provided clearly (σ_0) or with different degrees of blur (σ_M and σ_L), or withheld (σ_∞). Subjects are required to place the cursor on the target, thereby compensating for the lateral shift. The finger paths illustrate typical trajectories at the end of the experiment when the lateral shift was 2 cm (the colours correspond to two of the feedback conditions). **b**, The experimentally imposed prior distribution of lateral shifts is gaussian with a mean of 1 cm. **c**, A diagram of the probability distribution of possible visually experienced shifts under the clear and the two blurred feedback conditions (colours as in **a**) for a trial in which the true lateral shift is 2 cm. **d**, The estimate of the lateral shift for an optimal observer that combines the prior with the evidence. **e**, The average lateral deviation from the target as a function of the true lateral shift for the models (for σ_L the green shading shows the variability of the lateral deviation). Left, the full compensation model; middle, the bayesian probabilistic model; right, the mapping model (see the text for details).

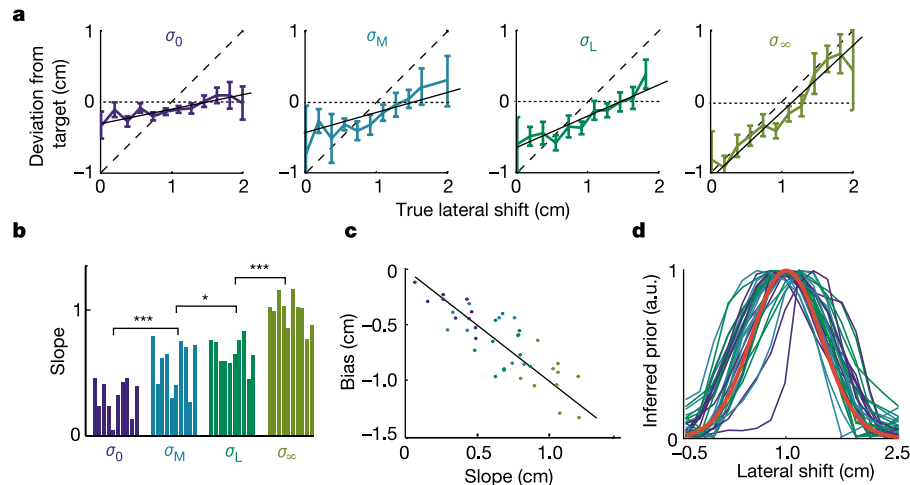


Figure 2 Results for a gaussian distribution. Colour codes as in Fig. 1. **a**, The lateral deviation of the cursor at the end of the trial as a function of the imposed lateral shift for a typical subject. Error bars denote s.e.m. The horizontal dotted lines indicate the prediction from the full compensation model and the dashed line is the fit for a model that ignores sensory feedback on the current trial and corrects only for the mean over all trials. The solid line is the bayesian model with the level of uncertainty fitted to the data. **b**, The slopes for the linear fits are shown for the full population of subjects. On the basis of the

hypothesis that the slope should increase with increasing visual uncertainty, we performed a repeated-measures analysis of variance on the slope, with visual uncertainty as a factor (main effect of visual uncertainty $F_{3,27} = 82.7$; $p < 0.001$). Planned comparisons of the slopes between adjacent uncertainty levels were all significant (asterisk, $p < 0.05$; three asterisks, $p < 0.001$). **c**, The bias against gain for the linear fits for each subjects and condition. The solid line shows the bayesian solutions. **d**, The inferred priors and the true prior (red) for each subject and condition.

visual estimate of the lateral shift. In this model, increasing the uncertainty of the feedback for a particular lateral shift (by increasing the blur) would affect the variability of the pointing but not the average location. Crucially, this model does not require subjects to estimate their visual uncertainty or the prior distribution of shifts. Second (model 2), subjects could optimally use information about the prior distribution and the uncertainty of the visual feedback to estimate the lateral shift. We can see intuitively why model 1 is sub-optimal. If, on a given trial, the subject sensed a lateral shift of 2 cm, there are many true lateral shifts that can give rise to such a perception. For example, the true lateral shift could be 1.8 cm with a visual error of +0.2 cm, or it could be a lateral shift of 2.2 cm with a visual error of -0.2 cm. Which of the two possibilities is more probable? Given gaussian noise on the visual feedback, visual errors of +0.2 cm and -0.2 cm are equally probable. However, a true lateral shift of 1.8 cm is more probable than a shift of 2.2 cm given that the prior distribution has a mean of 1 cm (Fig. 1b). If we consider all possible shifts and visual errors that can give rise to a sensed shift of 2 cm, we find that the most probable true shift is less than 2 cm. The amount by which it is less depends on two factors, the prior distribution and the degree of uncertainty in the visual feedback. As we increase the blur, and thus the degree of uncertainty, the estimate moves away from the visually sensed shift towards the mean of the prior distribution (Fig. 1d). Without any feedback (σ_∞) the estimate should be the mean of the prior. Such a strategy can be derived from bayesian statistics and minimizes the subject's mean squared error.

A third computational strategy (model 3) is to learn a mapping from the visual feedback to an estimate of the lateral shift. By minimizing the error over repeated trials, subjects could achieve a combination similar to model 2 but without any explicit representation of the prior distribution or visual uncertainty. However, to learn such a mapping requires knowledge of the error at the end of the movement. In our experiment we only revealed the shifted position of the finger at the end of the movement on the clear feedback trials (σ_0). Therefore, if subjects learn a mapping, they can only do so for these trials and apply the same mapping to the blurred conditions (σ_M , σ_L). This model therefore predicts that the average shift of the response towards the mean of the prior should be the same for all amounts of blur.

By examining the influence of the visual feedback on the final deviation from the target we can distinguish between these three models (Fig. 1e). If subjects compensate fully for the visual feedback (model 1), the average lateral deviation of the cursor from the target should be zero for all conditions. If subjects combine the prior and the evidence provided by sensory feedback (model 2), the estimated lateral shift should move towards the mean of the prior by an amount that depends on the sensory uncertainty. For a gaussian distribution of sensory uncertainty, this predicts a linear relationship between lateral deviation and the true lateral shift, which should intercept the abscissa at the mean of the prior (1 cm) and with a slope that increases with uncertainty. Finally, the mapping model (model 3) predicts that subjects should compensate for the seen position independently of the degree of uncertainty. Thus, all conditions should exhibit the same slope as the clear feedback condition (σ_0) of model 2. An examination of the theoretically determined mean squared error for the three models shows that it is minimal for model 2. Even though model 1 is on average on target, the variability in the response is higher than in model 2 (green shading in Fig. 1e shows the variability for the σ_L condition), leading to a larger mean squared error.

The lateral deviation from the target as a function of the lateral shift is shown for a representative subject in Fig. 2a. This shows a slope that increases with increasing uncertainty and is, therefore, incompatible with models 1 and 3. As predicted by model 2, the influence of the feedback on the final pointing location decreases with increasing uncertainty. The slope increases significantly with uncertainty in the visual feedback over the subjects tested (Fig. 2b). The bias and the slope should have a fixed relationship if we assume that subjects do bayesian estimation. We expect no deviation from the target if the true lateral shift is at the mean of the prior (1 cm). This predicts that the sum of the slope and offset should be zero, as observed in Fig. 2c. Subjects thus combine prior knowledge of the distribution with sensory evidence to generate appropriate compensatory movements.

Assuming that subjects use a bayesian strategy, we can furthermore use the errors that the subjects made during the trials to infer their degree of uncertainty in the feedback. For the three levels of imposed uncertainty, σ_0 , σ_M and σ_L , we find that subjects' estimates of their visual uncertainty are 0.36 ± 0.04 , 0.67 ± 0.1 and

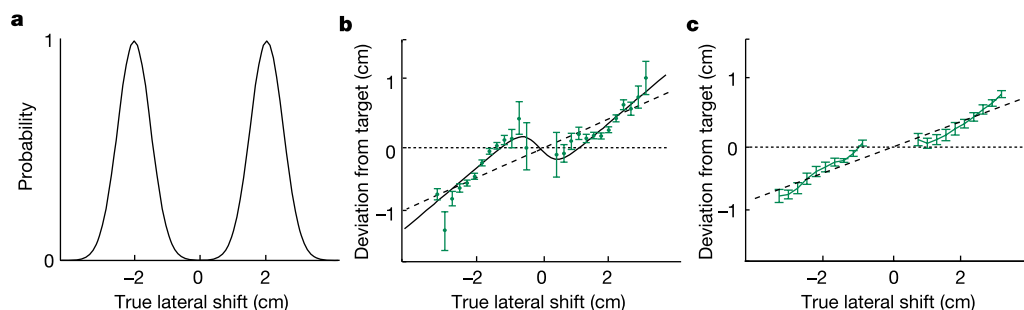


Figure 3 Results for a mixture of gaussian distributions. **a**, The experimentally imposed prior distribution of lateral shifts is a mixture of two gaussians. **b**, The lateral deviation of the cursor at the end of the trial as a function of the true lateral shift for a typical subject. Error bars denote s.e.m. The horizontal dotted lines indicate the prediction from

the full compensation model, the dashed line is the fit for a bayesian model with a single gaussian prior, and the solid line is the fit for a bayesian model with a prior that is a mixture of two gaussians. **c**, The lateral deviation across subjects (mean $\pm$ s.e.m. across subjects) is shown with a linear regression fit, demonstrating the nonlinearity of the data.

0.8 ± 0.1 cm (means $\pm$ s.e.m. across subjects), respectively. We have also developed a novel technique that uses these estimates to infer the priors used by the subjects. Figure 2d shows the priors inferred for each subject and condition. This shows that the true prior (red line) was reliably learned by each subject.

To examine whether subjects can learn complex distributions, a new group of subjects were exposed to a bimodal distribution (Fig. 3a) consisting of a mixture of two gaussians separated by 4 cm. Here, the bayesian model predicts a nonlinear relationship between true shift and lateral deviation, with the precise shape depending on the uncertainty of the visual feedback. Figure 3b shows a single subject's lateral deviation together with the fit of a bayesian model (solid line) in which we fit two parameters: the separation of the two gaussians and the variance of the visual uncertainty. The nonlinear properties are reflected in the empirical data and are consistent over the subjects (Fig. 3c) with a fitted separation of 4.8 ± 0.8 cm (mean $\pm$ s.e.m. across subjects), close to the true value of 4 cm, suggesting that subjects represent the bimodal prior. Taken together, our results demonstrate that subjects implicitly use bayesian statistics.

Many technically challenging problems have been addressed successfully within the bayesian framework^{7,8}. It has been proposed that the architecture of the nervous system is well suited for bayesian inference^{9–13} and that some visual illusions can be understood within the bayesian framework¹⁴. However, most models of the sensorimotor system consider a cascade of mappings from sensory inputs to the motor output^{15–17}. These models consider input–output relationships and do not explicitly take into account the probabilistic nature of either the sensors or the task. Recent models of motor control have begun to emphasize probabilistic properties^{18–24}. Unlike the visual system, which loses much of its plasticity once it has passed its critical period, the motor system retains much of its plasticity throughout adult life. We could therefore impose a novel prior on the subjects and measure its influence on sensorimotor processing. To show quantitatively that the system performs optimally would require a direct measure of sensory uncertainty before it is integrated with the prior. However, such a measure cannot easily be obtained as even a naive subject would integrate feedback with their natural, but unknown, prior. However, by imposing experimentally controlled priors we have shown that our results qualitatively match a bayesian integration process. A bayesian view of sensorimotor learning is consistent with neurophysiological studies showing that the brain represents the degree of uncertainty when estimating rewards^{25–27} and with psychophysical studies addressing the timing of movements^{28,29}. Although we have shown only the use of a prior in learning hand trajectories during a visuomotor displacement, we expect that such a bayesian process might be fundamental to all aspects of sensorimotor control and learning. For example, representing the distribution of dynamics of

objects, such as their mass, would facilitate our interactions with them. Similarly, although the possible configurations of the human body are immense, they are not all equally likely and knowledge of their distribution could be used to refine estimates of our current state. Taking into account a priori knowledge might be key to winning a tennis match. Tennis professionals spend a great deal of time studying their opponent before playing an important match, ensuring that they start the match with correct a priori knowledge. □

Methods

Experimental details

Six male and four female subjects participated in this study after giving informed consent. Subjects made reaching movements on a table during which an Optotrak 3020 tracking system (Northern Digital) measured the position of their right index finger. A projection–mirror system prevented direct view of their arm and allowed us to generate a cursor representing their finger position that could be displayed in the plane of the movement (for details of the set-up see ref. 30). Subjects saw a blue sphere representing the starting location, a green sphere representing the target and a white sphere representing the position of their finger (Fig. 1a). Subjects were requested to point accurately to the target. When the finger left the start position, the cursor representing the finger was extinguished and displaced to the right by an amount that was drawn each trial from a gaussian distribution with mean of 1 cm and standard deviation of 0.5 cm. Midway through the movement (10 cm), feedback of the cursor centred at the displaced finger position was flashed for 100 ms. On each trial one of four types of feedback (σ_0 , σ_M , σ_L , σ_∞) was displayed; the selection of the feedback was random, with the relative frequencies of the four types being (3, 1, 1, 1) respectively. The σ_0 feedback was a small white sphere. The σ_M feedback was 25 small translucent spheres, distributed as a two-dimensional gaussian with a standard deviation of 1 cm, giving a cloud-type impression. The σ_L feedback was analogous but had a standard deviation of 2 cm. No feedback was provided in the σ_∞ case. After another 10 cm of movement the trial was finished; feedback of the final cursor location was provided only in the σ_0 condition. The experiment consisted of 2,000 trials for each subject. On post-experimental questioning, all subjects reported being unaware of the displacement of the visual feedback. Only the last 1,000 trials were used for analysis. Subjects were instructed to take into account what they saw at the midpoint and to get as close to the target as possible; we took the lateral deviation of the finger from the target as a measure of subjects' estimate of the lateral shift. By averaging over trials we could obtain this estimate uncorrupted by any motor output noise, which we assumed to have mean of zero.

Bayesian estimation

We wish to estimate the lateral shift x_{true} of the current trial given a sensed shift x_{sensed} (also known as the evidence) and the prior distribution of lateral shifts $p(x_{\text{true}})$. From Bayes rule we can obtain the posterior distribution, that is the probability of each possible lateral shift taking into account both the prior and the evidence,

$$p(x_{\text{true}}|x_{\text{sensed}}) = p(x_{\text{sensed}}|x_{\text{true}}) \frac{p(x_{\text{true}})}{p(x_{\text{sensed}})}$$

where $p(x_{\text{sensed}}|x_{\text{true}})$ is the likelihood of perceiving x_{sensed} when the lateral shift really is x_{true} . We assume that visual estimation is unbiased and corrupted by gaussian noise so that

$$p(x_{\text{true}}|x_{\text{sensed}}) = \underbrace{\frac{1}{\sigma_{\text{sensed}} \sqrt{2\pi}} e^{-(x_{\text{true}} - x_{\text{sensed}})^2 / 2\sigma_{\text{sensed}}^2}}_{p(x_{\text{sensed}}|x_{\text{true}})} \underbrace{\frac{1}{\sigma_{\text{prior}} \sqrt{2\pi}} e^{-(x_{\text{true}} - 1\text{cm})^2 / 2\sigma_{\text{prior}}^2}}_{p(x_{\text{true}})} \underbrace{p(x_{\text{sensed}})}_{p(x_{\text{sensed}})}$$

For the optimal estimate we can find the maximum by differentiation, which represents the most probable lateral shift. For gaussian distributions such an estimate also has the

smallest mean squared error. This estimate is a weighted sum of the mean of the prior and the sensed feedback position:

$$x_{\text{estimated}} = \frac{\sigma_{\text{sensed}}^2}{\sigma_{\text{sensed}}^2 + \sigma_{\text{prior}}^2} [1 \text{ cm}] + \frac{\sigma_{\text{prior}}^2}{\sigma_{\text{sensed}}^2 + \sigma_{\text{prior}}^2} x_{\text{sensed}}$$

Given that we know σ_{prior}^2 , we can estimate the uncertainty in the feedback σ_{sensed} by linear regression from Fig. 2a.

Resulting mean squared error

The mean squared error (MSE) is determined by integrating the squared error over all possible sensed feedbacks and actual lateral shifts

$$\text{MSE} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_{\text{estimated}} - x_{\text{true}})^2 p(x_{\text{sensed}} | x_{\text{true}}) p(x_{\text{true}}) dx_{\text{sensed}} dx_{\text{true}}$$

For model 1, $x_{\text{estimated}} = x_{\text{sensed}}$, and this gives $\text{MSE} = \sigma_{\text{sensed}}^2$.

Using the result for $x_{\text{estimated}}$ from above for model 2 gives $\text{MSE} = \sigma_{\text{sensed}}^2 \sigma_{\text{prior}}^2 / (\sigma_{\text{sensed}}^2 + \sigma_{\text{prior}}^2)$, which is always lower than the MSE for model 1. If the variance of the prior is equal to the variance of the feedback, the MSE for model 2 is half that of model 1.

Inferring the used prior

An obvious choice of $x_{\text{estimated}}$ is the maximum of the posterior

$$p(x_{\text{true}} | x_{\text{sensed}}) = \frac{1}{\sigma_{\text{sensed}} \sqrt{2\pi}} e^{-(x_{\text{true}} - x_{\text{sensed}})^2 / 2\sigma_{\text{sensed}}^2} p(x_{\text{true}}) / p(x_{\text{sensed}})$$

The derivative of this posterior with respect to x_{true} must vanish at $x_{\text{estimated}}$. This allows us to estimate the prior used by each subject. Differentiating and setting to zero we get

$$\left. \frac{dp(x_{\text{true}})}{dx_{\text{true}}} \right|_{x_{\text{estimated}}} = \frac{(x_{\text{estimated}} - x_{\text{sensed}})}{\sigma_{\text{sensed}}^2}$$

We assume that x_{sensed} has a narrow peak around x_{true} and thus approximate it by x_{true} . We insert the σ_{sensed} obtained above, affecting the scaling of the integral but not its form. The average of x_{sensed} across many trials is the imposed shift x_{true} . The right-hand side is therefore measured in the experiment and the left-hand side approximates the derivative of $\log(p(x_{\text{true}}))$. Since $p(x_{\text{true}})$ must approach zero for both very small and very large x_{true} , we subtract the mean of the right-hand side before integrating numerically to obtain $\log(p(x_{\text{true}}))$, which we can then transform to estimate the prior $p(x_{\text{true}})$.

Bimodal distribution

Six new subjects participated in a similar experiment in which the lateral shift was bimodally distributed as a mixture of two gaussians:

$$p(x_{\text{true}}) = \frac{1}{2\sqrt{2\pi}\sigma_{\text{prior}}} \left(e^{-(x - x_{\text{sep}}/2)^2 / \sigma_{\text{prior}}^2} + e^{-(x + x_{\text{sep}}/2)^2 / \sigma_{\text{prior}}^2} \right)$$

where $x_{\text{sep}} = 4$ cm and $\sigma_{\text{prior}} = 0.5$ cm. Because we expected this prior to be more difficult to learn, each subject performed 4,000 trials split between two consecutive days. In addition, to speed up learning, feedback midway through the movement was always blurred (25 spheres distributed as a two-dimensional gaussian with a standard deviation of 4 cm), and feedback at the end of the movement was provided on every trial. Fitting the bayesian model (using the correct form of the prior and true σ_{prior}) to minimize the MSE between actual and predicted lateral deviations of the last 1,000 trials was used to infer the subject's internal estimates of both x_{sep} and σ_{sensed} . Some aspects of the nonlinear relationship between lateral shift and lateral deviation (Fig. 3a) can be understood intuitively. When the sensed shift is zero, the actual shift is equally likely to be to the right or the left and, on average, there should be no deviation from the target. If the sensed shift is slightly to the right, such as at 0.25 cm, then the actual shift is more likely to come from the right-hand gaussian than the left, and subjects should point to the right of the target. However, if the sensed shift is far to the right, such as at 3 cm, then because the bulk of the prior lies to the left, subjects should point to the left of the target.

Received 30 June; accepted 10 October 2003; doi:10.1038/nature02169.

- van Beers, R. J., Sittig, A. C. & Gon, J. J. Integration of proprioceptive and visual position-information: An experimentally supported model. *J. Neurophysiol.* **81**, 1355–1364 (1999).
- Jacobs, R. A. Optimal integration of texture and motion cues to depth. *Vision Res.* **39**, 3621–3629 (1999).
- Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
- Hillis, J. M., Ernst, M. O., Banks, M. S. & Landy, M. S. Combining sensory information: mandatory fusion within, but not between, senses. *Science* **298**, 1627–1630 (2002).
- Cox, R. T. Probability, frequency and reasonable expectation. *Am. J. Phys.* **17**, 1–13 (1946).
- Bernardo, J. M. & Smith, A. F. M. *Bayesian Theory* (Wiley, New York, 1994).
- Berrou, C., Glavieux, A. & Thitimajshima, P. Near Shannon limit error-correcting coding and decoding: turbo-codes. *Proc. ICC'93, Geneva, Switzerland* 1064–1070 (1993).
- Simoncelli, E. P. & Adelson, E. H. Noise removal via Bayesian wavelet coring. *Proc. 3rd International Conference on Image Processing, Lausanne, Switzerland, September*, 379–382 (1996).
- Olshausen, B. A. & Millman, K. J. in *Advances in Neural Information Processing Systems* vol. 12 (eds Solla, S. A., Leen, T. K. & Muller, K. R.) 841–847 (MIT Press, 2000).
- Rao, R. P. N. An optimal estimation approach to visual perception and learning. *Vision Res.* **39**, 1963–1989 (1999).
- Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. The 'wake-sleep' algorithm for unsupervised neural networks. *Science* **268**, 1158–1161 (1995).

- Sahani, M. & Dayan, P. Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Comput.* **15**, 2255–2279 (2003).
- Yu, A. J. & Dayan, P. Acetylcholine in cortical inference. *Neural Netw.* **15**, 719–730 (2002).
- Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nature Neurosci.* **5**, 598–604 (2002).
- Soechting, J. F. & Flanders, M. Errors in pointing are due to approximations in sensorimotor transformations. *J. Neurophysiol.* **62**, 595–608 (1989).
- Krakauer, J. W., Pine, Z. M., Ghilardi, M. F. & Ghez, C. Learning of visuomotor transformations for vectorial planning of reaching trajectories. *J. Neurosci.* **20**, 8916–8924 (2000).
- Lacquaniti, F. & Caminiti, R. Visuo-motor transformations for arm reaching. *Eur. J. Neurosci.* **10**, 195–203 (1998).
- Van Beers, R. J., Baraduc, P. & Wolpert, D. M. Role of uncertainty in sensorimotor control. *Phil. Trans. R. Soc. Lond. B* **357**, 1137–1145 (2002).
- van Beers, R. J., Wolpert, D. M. & Haggard, P. When feeling is more important than seeing in sensorimotor adaptation. *Curr. Biol.* **12**, 834–837 (2002).
- Harris, C. M. & Wolpert, D. M. Signal-dependent noise determines motor planning. *Nature* **394**, 780–784 (1998).
- Todorov, E. & Jordan, M. I. Optimal feedback control as a theory of motor coordination. *Nature Neurosci.* **5**, 1226–1235 (2002).
- Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. An internal model for sensorimotor integration. *Science* **269**, 1880–1882 (1995).
- Vetter, P. & Wolpert, D. M. Context estimation for sensorimotor control. *J. Neurophysiol.* **84**, 1026–1034 (2000).
- Scheidt, R. A., Dingwell, J. B. & Mussa-Ivaldi, F. A. Learning to move amid uncertainty. *J. Neurophysiol.* **86**, 971–985 (2001).
- Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–1902 (2003).
- Basso, M. A. & Wurtz, R. H. Modulation of neuronal activity in superior colliculus by changes in target probability. *J. Neurosci.* **18**, 7519–7534 (1998).
- Platt, M. L. & Glimcher, P. W. Neural correlates of decision variables in parietal cortex. *Nature* **400**, 233–238 (1999).
- Carpenter, R. H. & Williams, M. L. Neural computation of log likelihood in control of saccadic eye movements. *Nature* **377**, 59–62 (1995).
- Gold, J. I. & Shadlen, M. N. The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *J. Neurosci.* **23**, 632–651 (2003).
- Goodbody, S. J. & Wolpert, D. M. Temporal and amplitude generalization in motor learning. *J. Neurophysiol.* **79**, 1825–1838 (1998).

Acknowledgements We thank Z. Ghahramani for discussions, and J. Ingram for technical support. This work was supported by the Wellcome Trust, the McDonnell Foundation and the Human Frontiers Science Programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.P.K. (e-mail: konrad@koering.de).

Functional genomic hypothesis generation and experimentation by a robot scientist

Ross D. King¹, Kenneth E. Whelan¹, Ffion M. Jones¹, Philip G. K. Reiser¹, Christopher H. Bryant², Stephen H. Muggleton³, Douglas B. Kell⁴ & Stephen G. Oliver⁵

¹Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK

²School of Computing, The Robert Gordon University, Aberdeen AB10 1FR, UK

³Department of Computing, Imperial College, London SW7 2AZ, UK

⁴Department of Chemistry, UMIST, P.O. Box 88, Manchester M60 1QD, UK

⁵School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Manchester M13 9PT, UK

The question of whether it is possible to automate the scientific process is of both great theoretical interest^{1,2} and increasing practical importance because, in many scientific areas, data are being generated much faster than they can be effectively analysed. We describe a physically implemented robotic system that applies techniques from artificial intelligence^{3–8} to carry out cycles of scientific experimentation. The system automatically

originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using a laboratory robot, interprets the results to falsify hypotheses inconsistent with the data, and then repeats the cycle. Here we apply the system to the determination of gene function using deletion mutants of yeast (*Saccharomyces cerevisiae*) and auxotrophic growth experiments⁹. We built and tested a detailed logical model (involving genes, proteins and metabolites) of the aromatic amino acid synthesis pathway. In biological experiments that automatically reconstruct parts of this model, we show that an intelligent experiment selection strategy is competitive with human performance and significantly outperforms, with a cost decrease of 3-fold and 100-fold (respectively), both cheapest and random-experiment selection.

The branch of artificial intelligence devoted to developing algorithms for acquiring scientific knowledge is known as 'scientific discovery'. The pioneering work in the field was the development of learning algorithms for analysis of mass-spectrometric data³. In the subsequent 30 years much has been achieved and there are now several convincing examples in which computer programs have made explicit contributions to scientific knowledge^{4–8}. However, the general impact of such programs on science has been limited. This could now change because the expansion of automation in science is making it increasingly possible to couple scientific discovery software to laboratory instrumentation⁵. Here we describe a novel system that extends the existing level of integration of scientific discovery and robotics: the 'Robot Scientist' (Fig. 1).

A widely accepted view of science is that it follows a 'hypothetico-deductive' process¹. Scientific expertise and imagination are first used to form possible hypotheses, and then the deductive consequences of these hypotheses are tested by experiment. The Robot Scientist follows this paradigm: we employ the logical inference mechanism of abduction¹⁰ to form new hypotheses, and that of deduction to test which hypotheses are consistent (see Methods and Supplementary Information). The system has been physically implemented and conducts biological assays with minimal human intervention after the robot is set up. The hardware platform consists of a liquid-handling robot with its control PC, a plate reader with its control PC, and a master PC to control the system and do the scientific reasoning. The software platform consists of background knowledge about the biological problem, a logical inference engine, hypothesis generation code (abduction), experiment selection code (deduction), and the Laboratory Information Management System (LIMS) code that integrates the whole system. The robot conducts assays by pipetting and mixing liquids on microtitre plates. Given a computed definition of one or more experiments, we have developed code that designs a layout of reagents on the liquid-handling platform that will allow these experiments, with controls, to be performed efficiently. In addition, the liquid-handling robot is automatically programmed to plate out the yeast and media into the correct wells. The system measures the concentration of yeast in the wells of the microtitre trays using the adjacent plate reader and returns the results to the LIMS (although microtitre trays are still moved in and out of incubators manually).

The key point is that there was no human intellectual input in the design of experiments or the interpretation of data.

To test the Robot Scientist we chose the field of functional genomics. It is one of the most important in contemporary science and is an area in which laboratory automation is already mature. The current state of the art in functional genomics is to use highly automated robotics to generate data, and then to use data-mining systems to extract knowledge from that data. Within functional genomics, we selected yeast (*S. cerevisiae*) to test the methodology. Despite being one of the best understood of organisms, the functions of about 30% of the 6,000 genes in yeast are still unknown¹¹. The aim was to develop a system that could automatically determine the function of genes from the performance of knockout mutants (strains in which one gene has been removed). We focused on the aromatic amino acid synthesis (AAA) pathway (Fig. 2), and used auxotrophic growth experiments to assess the behaviour (phenotype) of the mutants. The AAA pathway is relatively well understood and of sufficient complexity to make reasoning about it non-trivial, and its intermediary metabolites are commercially available. Auxotrophic growth experiments consist of growing auxotrophic mutants on chemically defined media (a defined base plus one or more intermediate or terminal metabolites in the pathway), and observing whether growth is recovered or not (see Supplementary Information for details). A knockout mutant is auxotrophic if it cannot grow on a defined medium on which the wild type can grow. Auxotrophic experiments are a classic technique for inferring metabolic pathways⁹.

We needed a way of representing prior biological knowledge in the computer, so we developed a logical formalism to model cellular metabolism that captures the key relationship between protein-coding sequences (open reading frames; ORFs), enzymes and metabolites in a pathway¹². All objects (ORFs, proteins and metabolites) and relationships (coding, reactions, transport and feedback) are described as logical formulae. The structure of the metabolic pathway is that of a directed graph, with metabolites as nodes and enzymes as arcs. An arc corresponds to a reaction. The compounds at each node are the set of all metabolites that can be synthesized by the reactions leading to it. Reactions are modelled as unidirectional transformations. Using this formalism, we have implemented a model of the AAA pathway with the logic programming language Prolog (the complete model is provided in Supplementary Information). Prolog makes it possible both to inspect the biological knowledge in the model directly and to compute the predictions of the model automatically. The model infers (deduces) that a knockout mutant will grow if, and only if, a path can be found from the input metabolites to the three aromatic amino acids. This allows the model to compute the phenotype of a particular knockout or to be used to infer missing reactions that could explain an observed phenotype (abduction). We consider that most hypothesis generation in modern biology is abductive. What is inferred are not general hypotheses, which would be inductive, but specific facts about biological entities.

The original bioinformatic information for the AAA model was taken mainly from the KEGG¹³ catalogue of metabolism. The model was then tested with all possible auxotrophic experiments involving a single replacement metabolite, and was altered manually to fit the empirical results. To ensure that the model was not 'over-fitted', we carried out all possible auxotrophic experiments with pairs of metabolites. The model correctly predicted at least 98.5% of the experiments (Supplementary Information). To the best of our knowledge, no bioinformatic model has been as thoroughly tested with knockout mutants.

Machine learning is the branch of artificial intelligence that seeks to develop computer systems that improve their performance automatically with experience^{14,15}. It has much in common with statistics, but differs in having a greater emphasis on algorithms, data representation and making acquired knowledge explicit. The

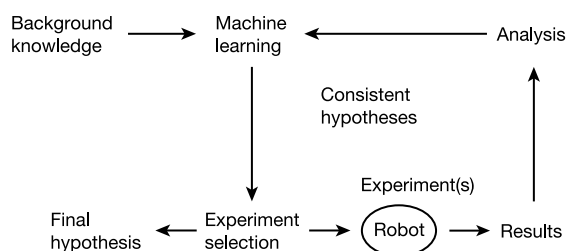


Figure 1 The Robot Scientist hypothesis-generation and experimentation loop.

branch of machine learning that deals with algorithms that can choose experiments is known as ‘active learning’^{16,17}. If we assume that each hypothesis has a prior probability of being correct and that each experiment has an associated cost, then scientific experiment selection can be formalized as the task of selecting the optimal series of experiments (in terms of expected cost) to eliminate all except the one correct hypothesis. This problem is, in general, computationally intractable (NP-hard)¹⁸. However, it can be shown that the problem is structurally the same as finding the smallest decision tree—experiments are nodes and hypotheses are leaves¹⁹. This is significant because a bayesian analysis of decision-tree learning has shown that near-optimal solutions can be found in polynomial time¹⁹. This analysis leads to the following approximate recurrence formula for the expected cost²⁰. Let $EC(H, T)$ denote the minimum expected cost of experimentation given the set of candidate hypotheses H and the set of candidate trials T :

$$EC(\emptyset, T) = 0$$

$$EC(\{h\}, T) = 0$$

$$EC(H, T) \approx \min_{t \in T} [C_t + p(t)(\text{mean}_{t' \in (T-t)} C_{t'}) J_{H[t]} + (1 - p(t))$$

$$\times (\text{mean}_{t' \in (T-t)} C_{t'}) J_{H[t]}]$$

$$J_H = -\sum_{h \in H} P(h) [\log_2(p(h))]$$

where C_t is the monetary price of the trial t , $p(t)$ is the probability that the outcome of the trial t is positive, and [...] is the ‘floor’ function. $p(t)$ can be computed as the sum of the probabilities of the hypotheses (h) that are consistent with a positive outcome of t .

In current robot-scientist experiments, it is assumed that the system knows the biochemical equations of the AAA pathway in the wild type. What the system does not know is how this biochemistry is related to the genetics. The hypotheses are therefore bindings between the ORFs and enzymic reactions; that is, which of the 15 possible ORFs in the pathway has been deleted (although only 8 ORFs were ever actually used for technical reasons, the system was not aware of this). For example, a correct hypothesis would be that ORF YPR060c codes for chorismate mutase (which catalyses

the reaction chorismate $\rightarrow$ prephenate; Fig. 2). These hypotheses are automatically generated by abducting the different possibilities from the model^{12,20–22}. Hypotheses are evaluated by comparing their predicted consequences with the actual experimental results. These logical inferences are made with our machine learning system ASE-Progol²³, where ASE stands for ‘active selection of experiments’ (ASE-Progol is freely available to academics, from http://www.comp.rgu.ac.uk/pub/staff/chb/systems/ase_progol/version_1.0). We compared three experimental selection strategies: first, ASE (our approximation to choosing the experiment that minimizes the expected cost—see above); second, Naive (which chooses the cheapest experiment that has not yet been done); and last, Random (which chooses a random experiment that has not yet been done).

The performance of a strategy is defined as the average accuracy of all hypotheses not already eliminated by experiments using that strategy. The accuracy of a single hypothesis is the number of correct predictions that this hypothesis makes, for all possible single-metabolite and double-metabolite addition experiments, assuming that the model is 100% accurate. We evaluated accuracy versus the monetary price of the experiments (pounds sterling) and time (iterations; that is, the number of sequential experiments).

Using the Robot Scientist, we compared the experimental strategies ASE, Naive and Random over five experimental iterations (Fig. 3a, b). At the end of the fifth iteration, ASE had an estimated accuracy of 80.1%, in comparison with 74.0% for Naive and 72.2% for Random. At this point in the iterative cycle of experimentation, the ASE strategy was significantly more accurate than either Naive ($P < 0.05$) or Random ($P < 0.07$) using a paired t -test. Given a maximum spend of £10^{2,26} (the maximum that Naive spends; note that these prices are scaled), ASE has an estimated accuracy of 79.5%, in comparison with 73.9% for Naive and 57.4% for Random. For this price, ASE is significantly more accurate than either Naive ($P < 0.05$) or Random ($P < 0.001$). Although the AAA pathway is sufficiently complex to provide an adequate test for the robot scientist system, it is not so complex that the Naive and Random strategies will not eventually achieve a high accuracy. Where ASE shows its superiority over the other two experimental strategies is in

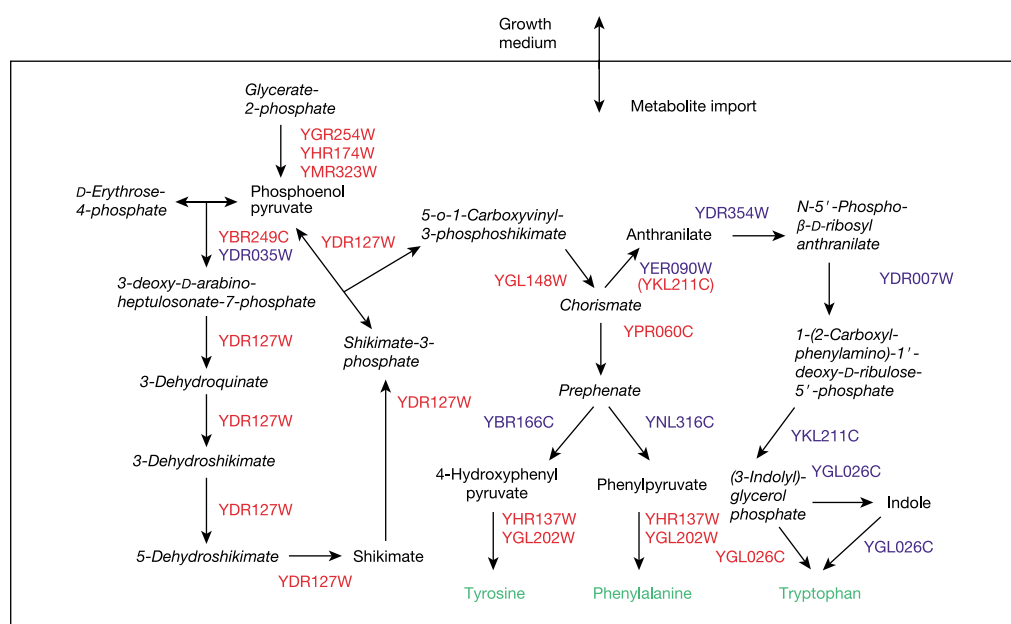


Figure 2 A schematic representation of the logical model of the aromatic amino acid pathway in yeast. The metabolites are the nodes and the enzymes the arcs in the directed graph. Metabolites that could not be used in auxotrophic experiments are in italics, and

the end-product aromatic amino acids are shown in green. The ORFs encoding the enzyme proteins are blue if they are amenable to auxotrophic experimentation; otherwise they are red. The logical model also represents the import of metabolites into the cell.

cost-effectiveness. In science, as in industry, 'time is money' (although the conversion rate may be unclear). The experimental cost we wish to optimize will therefore be a function of price and time. Because ASE dominates for both, ASE is superior to Naive and Random for any positive function of price and time. For example, to achieve an accuracy of about 70%, ASE requires fewer trial iterations, and a hundredth of the price, of Random, and almost half the number of iterations, and a third of the price, of Naive.

We evaluated the reliability of these empirical results by running several computer simulations of the Robot Scientist. In these we used the AAA Prolog model and the ASE-Progol inference system, and we took the results of manual experiments. When necessary, we also added random noise by flipping the observation of growth to no-growth and the reverse. We compared differing types of experiment (single metabolites or double metabolites), numbers of iterations, noise, and noise reduction strategies. In Fig. 3c, d we show the results for 0% and 25% experimental noise, plotting accuracy against time and accuracy against price. In these typical experiments, the ASE strategy almost completely dominates those of Naive and Random. These simulation experiments are consistent with the results *in vivo*, because we estimate that the physical system has about 25% noise in the observations of whether growth occurred or not. The high level of noise is due to the robot's being open to the air (and therefore to microbial contamination). We were also interested to compare the performance of the Robot Scientist system with that of humans. To do this, we adapted the simulator to allow humans to choose and interpret the result of cycles of experimentation. In initial trials, using nine graduate computer scientists and biologists, we found that there was no significant difference between the robot and the best human performance in terms of the number of iterations required to

achieve a given level of accuracy.

The inference of gene function in the AAA pathway was selected to be typical of scientific inference tasks in molecular genetics. We argue that most molecular genetics differs from this model only in the much greater sophistication of the abductive process used to select a possible hypothesis and by the far greater elegance of the experiments used to discriminate among hypotheses. Although the AAA problem is relatively simple, it could still be applied directly to 'real-world' problems. For instance, if a series of AAA metabolism mutants from a related (but genetically uncharacterized) fungal species had been isolated, the *S. cerevisiae* AAA pathway could be used to guide auxotrophic growth experiments in exactly the same way as described here.

Although pioneering work by others (see ref. 22, for example) has developed automated systems to induce genetic networks from mutant data, we believe that the Robot Scientist is the first example of a closed-loop system that actually designs and executes experiments to test inferred hypotheses. Nevertheless, the Robot Scientist has currently only been demonstrated to rediscover the role of genes of known function; this initial step was essential to ensure that the system was working. We now plan to extend the system to be able to uncover the function of genes whose role is currently unknown. To enable this, the background cellular metabolism model will need to be extended. This will involve the translation of bioinformatic databases such as KEGG into our logical formalism; the resultant models are likely to be less reliable than that of our carefully checked AAA model. The abductive hypothesis generation method will also need to be extended to allow the inference of missing arcs, not just their binding. Initial work has been done on this¹².

The general robot scientist idea could be applied to many scientific problems, and we are actively investigating drug design

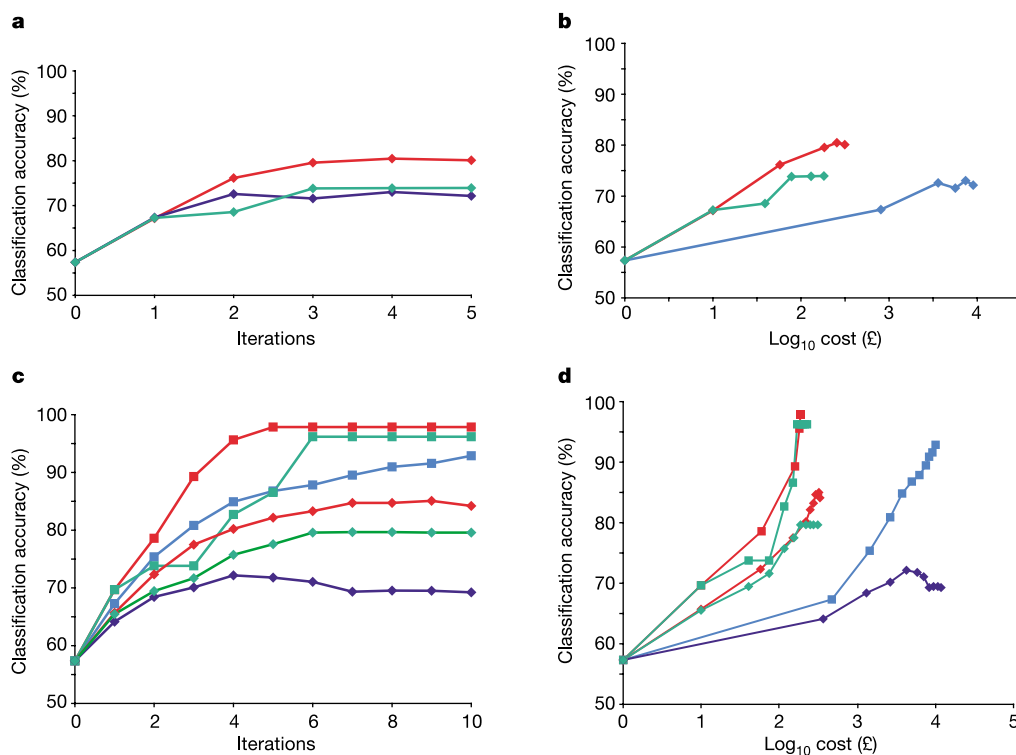


Figure 3 Actual and simulated performance of the Robot Scientist. **a, b**, Actual performance. **a**, Plot of average classification accuracy against experimental iteration (from four experimental repeats). **b**, Plot of average classification accuracy against average experimental scaled price (from four experimental repeats). **c, d**, Simulated performance. **c**, Plot of average classification accuracy against time for 0% noise

(squares) and 25% noise (diamonds) (based on 100 simulation runs). **d**, Plot of average classification accuracy against average scaled price for 0% noise (squares) and 25% noise (diamonds) (based on 100 simulation runs). Strategies: red curves, ASE; blue curves, Random; green curves, Naive.

and quantum control. To apply the Robot Scientist to drug design, the robot would need to be able to perform a biological assay to perform the given optimization, and either synthesize compounds of required structure using lab-on-a-chip technology or choose them from a large library. The hypotheses would be different qualitative structure–activity relationships. The quantum control of chemical synthesis with femtosecond lasers²⁴ is especially challenging because of the ability and desirability of doing experiments very quickly (about 10^3 per second at present). At such speeds, human intervention is impossible, and automatic systems become essential.

The Robot Scientist extends the state of the art in integrating scientific discovery software with laboratory robotics. Moreover, the application of the Robot Scientist to functional genomics provides further evidence that some aspects of scientific reasoning can be formalized and efficiently automated. In science, experiment selection is generally done informally and without explicit regard to the cost of eliminating hypotheses; when formal experimental design²⁵ is employed, this is generally done to ensure that sufficient repeats and controls are in place to answer the desired question. Our results show that different experiment selection strategies can have significantly different results in terms of cost, even for the solution of a simple problem. This suggests that there remains scope to improve the general cost-effectiveness of science by developing better tools to help choose efficient experiments.

Automation was the driving force of much of nineteenth-century and twentieth-century change, and this is likely to continue. It is also becoming increasingly important in scientific research: for example, the sequencing of the human genome was made possible by factory production techniques, and modern drug discovery relies on high-throughput screening robots. We consider this trend to increased automation of science to be both inevitable and desirable. It is inevitable because it will be required to deal with the challenges of science in the twenty-first century. It is also desirable because it frees scientists to make the high-level creative leaps at which they excel. □

Methods

Growth experiments and the amino acid pathway

The mutants (of *Saccharomyces cerevisiae* strain BY4741 (ATCC201388) *MatA his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*; ref. 26) had the complete reading frame of each protein-encoding gene deleted by replacement with a selectable marker gene that had no phenotype in the absence of the selective agent²⁷. They were therefore non-revertible null mutants. Limitations in the availability of mutant yeast strains restricted the number of genes used in the *in vivo* investigations to 15, and the auxotrophic experimental requirement for a difference in growth phenotype reduced this number to the following 8 (the other mutants always either grew or failed to grow): *ybr166c*, *ydr007w*, *ydr035w*, *ydr354w*, *yer090w*, *ygl026c*, *ykl211c* and *ynl316c*. The number of possible metabolites was limited to nine by availability and cost: anthranilate (10), indole (190), *p*-hydroxyphenolpyruvate (193), *L*-phenylalanine (53), phenylpyruvate (30), phosphoenolpyruvate (9385), shikimic acid (633), *L*-tyrosine (53) and *L*-tryptophan (53). The numbers in parentheses are the normalized true experimental costs of using each metabolite in a growth medium; note the approximate three orders of magnitude range.

Computational model of the aromatic amino acid pathway

The Prolog model of the pathway was refined in three stages. First, the original model was translated from KEGG¹³ and carefully checked with the literature. We checked that all the KEGG reactions were documented in *S. cerevisiae* (consistency) and that there were no other related reactions described in the literature (completeness). Second, the predictions of the model were compared with the results of the single-metabolite experiments (see above). Whether growth or lack of growth was observed was at this point decided visually. Because certain metabolites did not seem to affect growth in the way predicted by the literature, we refined the model to make these metabolites unable to be imported into the cells efficiently. It was also necessary to add inhibition effects (see Supplementary Information for details). Last, the results of the double-metabolite experiments were tested against the model, and the automatic growth-calling software was optimized by learning a decision tree to fit the experimental results to the model (see Supplementary section 2). The model developed on the single metabolites was consistent with all except less than 1.5% of the double-metabolite experiments. The model was not further changed to include these experimental discrepancies.

Performance measures

The average performance of the hypotheses is an appropriate performance measure

because it rewards learners that discriminate between competing hypotheses. This approach is a compromise between selecting the hypothesis with the highest probability and weighting all predictions by the probability of the hypotheses that generated them. In active learning, the performance curves that have been generally used plot predictive accuracy against the number of training examples. Often, two curves are plotted on the same graph, one for active learning and one for random sampling. The accuracy of a single hypothesis is the number of correct predictions that this hypothesis makes about all possible single-metabolite and double-metabolite experiments, based on using the model as the oracle.

Such performance plots allow the difference in the number of experiments (examples per unit time) required to reach a particular level of performance to be compared. However, one drawback of such plots is that they ignore any variation in the price of obtaining individual examples. When such variation does exist, and the aim is to compare the price of attaining particular levels of performance, these plots are potentially misleading. To overcome this drawback we also plot the cumulative price of the experiments against performance²⁰. For this we use the normalized price of the metabolite. At the start of the experiments, when there are eight possible hypotheses, the average accuracy is 57%.

Structure of the Robot Scientist

The Robot Scientist automates the task of liquid handling and can conduct assays by pipetting and mixing liquids on microtitre plates. The robot is controlled using Tcl (Tool Command Language), and we have written a compiler that translates Prolog commands into Tcl robot operations. Given a Prolog definition of one or more experiments, we have developed code that designs a layout of the robot that will allow these experiments, with controls, to be performed efficiently. In addition, the robot has to be automatically programmed to plate out the yeast and media into the correct wells. The microtitre plates were measured with the adjacent plate reader and the results were returned to the LIMS. Transfer of plates from the robot to the incubator, and from the incubator to the plate reader, was done manually.

Received 24 July; accepted 14 November 2003; doi:10.1038/nature02236.

1. Popper, K. *The Logic of Scientific Discovery* (Hutchinson, London, 1972).
2. Sloman, A. *The Computer Revolution in Philosophy* (Harvester, Hassocks, Sussex, 1978); available online from (<http://www.cs.bham.ac.uk/research/cogaff/crp/>).
3. Buchanan, B. G., Sutherland, G. L. & Feigenbaum, E. A. in *Machine Intelligence* Vol. 4 (eds Meltzer, B. & Michie, D.) 209–254 (Edinburgh Univ. Press, 1969).
4. Langley, P., Simon, H. A., Bradshaw, G. L. & Zytkow, J. M. *Scientific Discovery: Computational Explorations of the Creative Process* (MIT Press, Cambridge, Massachusetts, 1987).
5. Zytkow, J. M., Zhu, J. & Hussam, A. Automated discovery in a chemistry laboratory. in *Proceedings of the 8th National Conference on Artificial Intelligence (AAAI-1990)* (eds Dietterich, T. & Swartout, W.) 889–894 (MIT, Cambridge, Massachusetts, 1990).
6. King, R. D., Muggleton, S. H., Srinivasan, A. & Sternberg, M. J. E. Structure–activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proc. Natl Acad. Sci. USA* **93**, 438–442 (1996).
7. Valdes-Perez, R. E. Discovery tools for science applications. *Commun. ACM* **42**, 37–41 (1999).
8. Langley, P. The computational support of scientific discovery. *Int. J. Hum.–Comput. Stud.* **53**, 393–410 (2000).
9. Beadle, G. W. & Tatum, E. I. Genetic control of biochemical reactions in *Neurospora*. *Proc. Natl Acad. Sci. USA* **27**, 499–506 (1941).
10. Pierce, C. S. *Collected Papers of Charles Sanders Pierce* (Harvard Univ. Press, 1958).
11. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002); (<http://mips.gsf.de/proj/yeast/CYGD/db/index.html>).
12. Reiser, P. G. K. *et al.* Developing a logical model of yeast metabolism. *Electron. Trans. Artif. Intell.* **5**, 223–244 (2001).
13. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000); (<http://www.genome.ad.jp/kegg/>).
14. Mitchell, T. M. *Machine Learning* (McGraw-Hill, New York, 1997).
15. Langley, P. *Elements of Machine Learning* (Morgan Kaufmann, San Mateo, California, 1996).
16. Cohen, D. A., Ghabramani, Z. & Jordan, M. I. Active learning with statistical models. *J. Artif. Intell. Res.* **4**, 129–145 (1996).
17. Lin, F.-R. & Shaw, M. J. Active training of backpropagation neural networks using the learning by experimentation methodology. *Ann. Oper. Res.* **75**, 129–145 (1997).
18. Fedorov, V. V. *Theory of Optimal Experiments* (Academic, London, 1972).
19. Muggleton, S. & Page, D. in *Machine Intelligence* Vol. 15 (eds Furukawa, K., Michie, D. & Muggleton, S.) 248–267 (Oxford Univ. Press, 1999).
20. Bryant, C. H., Muggleton, S. H., Oliver, S. G., Kell, D. B., Reiser, P. G. K. & King, R. D. Combining inductive logic programming, active learning, and robotics to discover the function of genes. *Electron. Trans. Artif. Intell.* **5**, 1–36 (2001).
21. Flach, P. & Kaka, A. *Abduction and Induction* (Kluwer, London, 2000).
22. Zupan, B. *et al.* in *Proceedings of the Eighth European Conference on Artificial Intelligence in Medicine* (eds Qualini, S., Barahona, P. & Andreassen, S.) 304–313 (Springer, Berlin, 2001).
23. Muggleton, S. Inverse entailment and Prolog. *New Generation Comput. J.* **13**, 245–286 (1995).
24. Rabitz, H., de Vivie-Riedle, R., Motzkus, M. & Kompa, K. Whither the future of controlling quantum phenomena? *Science* **288**, 824–828 (2000).
25. Cochran, W. G. & Cox, G. M. *Experimental Designs* (Wiley, New York, 1992).
26. Brachmann, C. B. *et al.* Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**, 115–132 (1998).
27. Gaeffer, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank D. Page, U. Sarkans, A. Tamaddoni, M. Sternberg, A. Sloman and D. Michie for their help and advice, and D. Struttman for technical assistance. The work was funded by the BBSRC, the EPSRC, the Wellcome Trust and PharmDM.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to S.G.O. (steve.oliver@man.ac.uk).

Immune recognition of a human renal cancer antigen through post-translational protein splicing

Ken-ichi Hanada¹, Jonathan W. Yewdell² & James C. Yang¹

¹Surgery Branch, National Cancer Institute, National Institutes of Health, 9000 Rockville Pike, Building 10, Room 2B42, Bethesda, Maryland 20892, USA

²Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

Cytotoxic T lymphocytes (CTLs) detect and destroy cells displaying class I molecules of the major histocompatibility complex (MHC) that present oligopeptides derived from aberrant self or foreign proteins. Most class I peptide ligands are created from proteins that are degraded by proteasomes and transported, by the transporter associated with antigen processing, from the

cytosol into the endoplasmic reticulum, where peptides bind MHC class I molecules and are conveyed to the cell surface¹. C2 CTLs, cloned from human CTLs infiltrating a renal cell carcinoma, kill cancer cells overexpressing fibroblast growth factor-5 (FGF-5)². Here we show that C2 cells recognize human leukocyte antigen-A3 MHC class I molecules presenting a nine-residue FGF-5 peptide generated by protein splicing. This process, previously described strictly in plants³ and unicellular organisms⁴, entails post-translational excision of a polypeptide segment followed by ligation of the newly liberated carboxy-terminal and amino-terminal residues. The occurrence of protein splicing in vertebrates has important implications for the complexity of the vertebrate proteome and for the immune recognition of self and foreign peptides.

To identify the FGF-5-derived epitope recognized by C2, we transfected human leukocyte antigen (HLA)-A3-expressing COS-7 cells (COS-A3) with expression plasmids encoding truncated forms of the FGF-5 gene and measured C2 activation by its release of interferon- γ (IFN- γ) (Fig. 1a). Antigenicity was lost by truncation either between predicted amino acid residues 212 and 220 (G4/G5 plasmids) or residues 161 and 172 (G8/G9 plasmids) of FGF-5. The shortest stimulatory fragment identified (G8) encoded 60 amino acids (residues 161–220).

We next examined the antigenicity of synthetic peptides corresponding to all of the 8-residue, 9-residue and 10-residue peptides contained within the defined 60 amino-acid fragment, but none were recognized by C2 when incubated with APCs expressing HLA-A3 (data not shown). We similarly tested peptides encoded by the +1 and +2 reading frames of the fragment but did not find the epitope (data not shown). Post-translational structural modification of a minimal determinant within the 60 amino-acid fragment remained a possible explanation for our failure to identify a

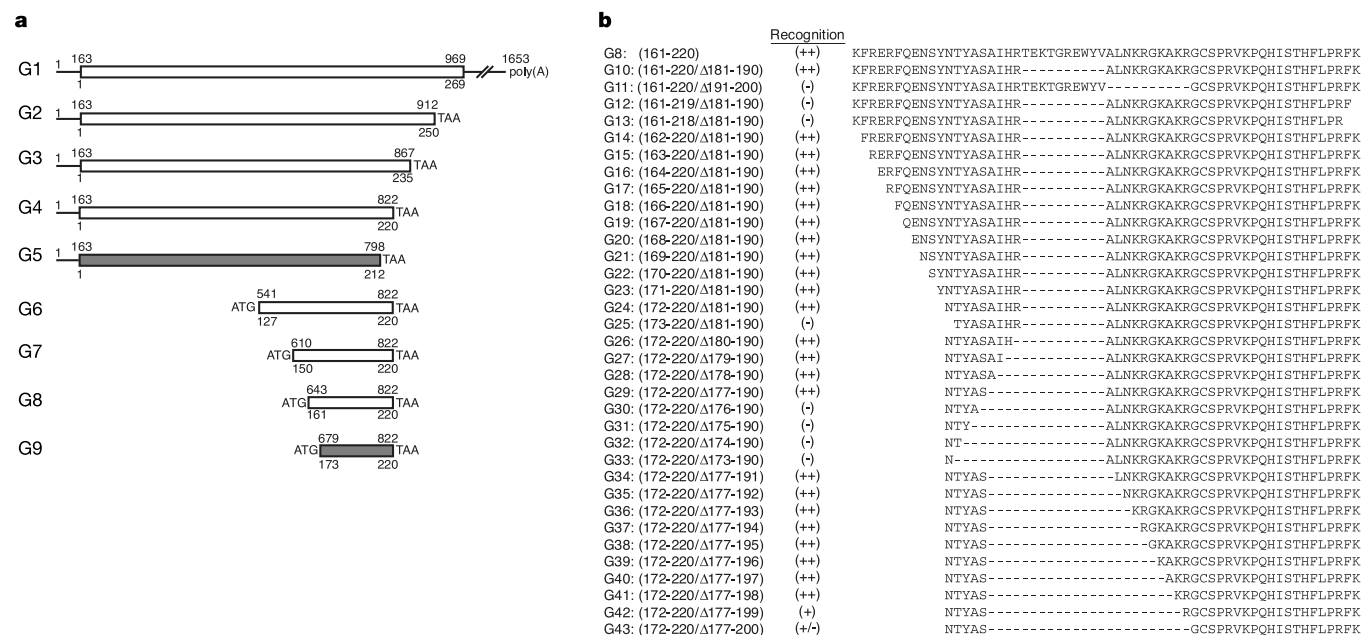


Figure 1 Genetic truncation analysis of the FGF-5-encoded determinant. **a**, A series of truncation mutants (G2–G9) were prepared by PCR from construct G1. An initiation codon and the Kozak consensus (CACCATG) were introduced into the PCR primers for constructs G6–G9. Each PCR product was cloned into the eukaryotic expression vector and transfected into COS-A3 cells; recognition by C2 was assessed by IFN- γ secretion. Nucleic acid positions are indicated above the start and finish of each construct and the corresponding amino acid positions from FGF-5 below. Constructs shown as open boxes were recognized; those shown as filled boxes were not. **b**, Analysis of internal deletion

mutants. For simplicity, predicted amino acid sequences instead of DNA sequences are shown. The antigenicity of each construct was assessed as described in **a**. All constructs were designed to start with the Kozak consensus sequence and end with the termination codon TAA. Numbers in parentheses indicate the position of the first and the last amino acids. Numbers after Δ indicate internally deleted amino acid residues. The degree of CTL recognition is categorized as follows depending on IFN- γ secretion: (++) , more than 1,000 pg ml⁻¹; (+) , 200–1,000 pg ml⁻¹; (+/-) , 100–200 pg ml⁻¹; (-) , less than 100 pg ml⁻¹.

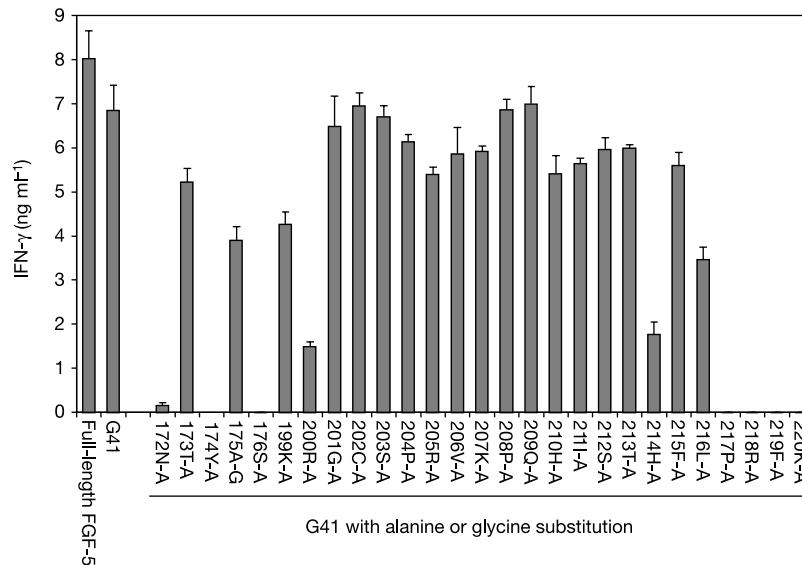


Figure 2 Effect of Ala or Gly substitution on the antigenicity of the G41 construct. Nucleotide substitutions encoding Ala were introduced into G41, and the constructs were examined for recognition as described in Fig. 1a. '172N-A', for example, indicates Ala

substitution for Asn at position 172. Position 175, which is naturally Ala, was replaced with Gly. Error bars show the standard deviation of duplicate IFN- γ determinations. Similar results were obtained more than twice.

synthetic peptide with antigenic activity. We therefore returned to a genetic approach to identify crucial residues within this fragment of FGF-5. Surprisingly, a plasmid (G10; Fig. 1b), internally deleted of amino acids 181–190, proved to be antigenic after transfection in COS-A3. A series of plasmids with more extensive deletions (C-terminal, G12/G13; N-terminal, G14–G25; internal N-terminal side, G26–G33; internal C-terminal side, G34–G43) were prepared and analysed. Plasmid G41, encoding FGF-5 residues 172–176 and 199–220, was the smallest construct that was recognized as strongly as full-length FGF-5. To identify the amino acids encoded by G41 that contribute to antigenicity, plasmids were constructed that substituted an Ala-encoding codon at each of the 27 codons within G41 (Fig. 2). This revealed that Ala substitutions in seven residues (Asn 172, Tyr 174, Ser 176, Pro 217, Arg 218, Phe 219 and Lys 220) abolish CTL recognition. In particular, peptides with Tyr at position 3 and Phe or Lys at positions 9 or 10 are favoured for binding to HLA-A3 molecules⁵, indicating that the epitope might be a peptide created from fusing separated genetic elements.

Analysing candidate synthetic 9-residue and 10-residue fusion peptides (Fig. 3a), we found that a nonameric peptide deriving from five N-terminal amino acids and four C-terminal amino acids (NTYASPRFK) was strongly recognized, stimulating C2 cells at a concentration of 10 nM. The only other stimulatory peptide was the decamer NTYASLPRFK, which demonstrated 1,000-fold lower activity (Fig. 3b). Covalent fusion was required between NTYAS and PRFK because incubating APCs with even high concentrations of these peptides failed to activate C2 cells (Fig. 3a, NTYAS + PRFK).

The generation of the nonameric fusion peptide from the FGF-5 gene and the various gene fragments we examined could result from RNA splicing, ribosome skipping⁶ or protein splicing⁴. RNA splicing is an unlikely explanation for two reasons: first, numerous internal truncation mutants recognized by C2 (Fig. 1b, constructs G26–G29 and G34–G41) possess different sequences flanking potential splice sites, and second, neither the original FGF-5 sequence nor any of the truncation mutants share splicing motifs⁷. To examine the possibility of ribosome skipping, we generated three plasmids with termination codons inserted between the two determinant-encoding fragments (Fig. 4a). No plasmids sensitized cells for C2 recognition after transfection into COS-A3 APCs (Fig. 4b).

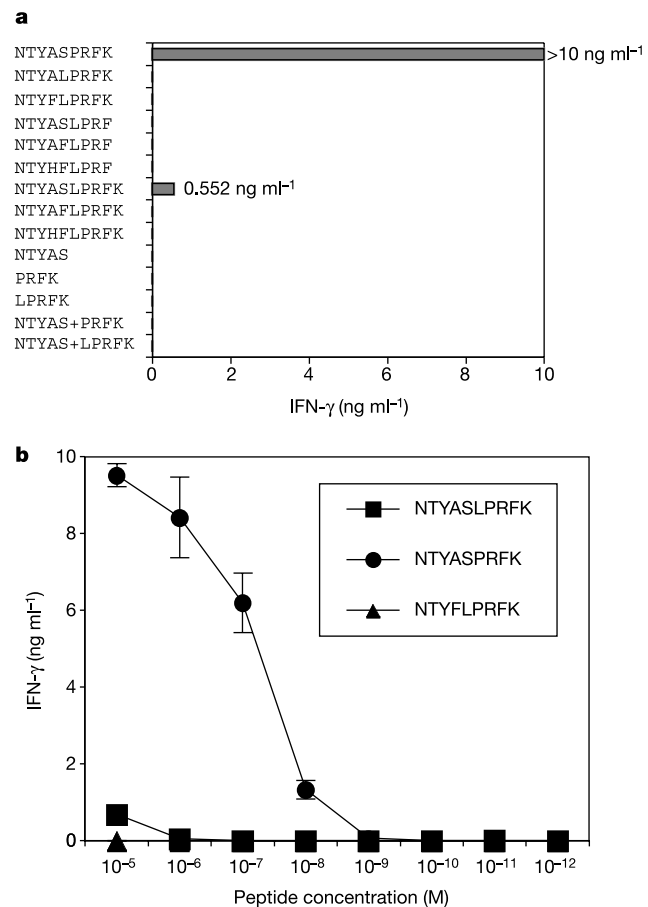


Figure 3 Identification of the nonamer. **a**, Peptides were pulsed onto autologous HLA-A3⁺ EBV-B cells at a concentration of 10 μ M; recognition by C2 was determined by IFN- γ release. None of the peptides were recognized when pulsed onto HLA-A3⁺ EBV-B cells (not shown). **b**, Titration of peptides. Three fusion peptides (NTYASPRFK, NTYASLPRFK and NTYFLPRFK) were pulsed onto the autologous EBV-B cell line at the concentrations indicated, and their recognition by C2 was assessed by IFN- γ assay. Error bars show the standard deviation of duplicate IFN- γ determinations. Similar results were obtained in at least two other independent experiments.

This also refutes epitope generation by aberrant RNA splicing because it seems that the intervening sequence between the two determinant-encoding fragments is transcribed and translated, indicating a post-translational mechanism.

To pursue this, we examined the processing and recognition of synthetic peptides from FGF-5 (Fig. 5a) pulsed onto HLA-A3⁺ or HLA-A3⁻ Epstein-Barr-virus-transformed B (EBV-B) cells (Fig. 5b, c). High concentrations of a peptide corresponding to FGF-5 172–220 starting with the 5-mer (172–176, NTYAS) and ending with the 4-mer (217–220, PRFK) activated C2 (but not another RCC-specific CTL clone) in a HLA-A3-restricted manner. Recognition was lost by the deletion of residue 172 or residues 213–220 and was not affected by the addition of residues 161–171. Two experiments eliminated the possibility of contamination of the 49-mer peptide with 9-mer fusion peptide. First, presentation of the long peptide was abrogated by mild aldehyde fixation of APCs, a process that did not affect the presentation of the 9-mer (Fig. 5d). Second, we fractionated synthetic 9-mer (NTYASPRFK) and 49-mer (172–220) peptides by high-performance liquid chromatography (HPLC) and tested the capacity of fractions to activate C2. The antigenic activities of the 9-mer and the 49-mer eluted in fractions 11 (Fig. 6a) and 22 (Fig. 6b), respectively. These fractions contained the expected peptides as determined by mass spectrometry (not shown). Again, the 9-mer eluting in fraction 11 was active with fresh or fixed APCs, while the 49-mer eluting in fraction 22 was active only with live cells.

We further used HPLC to demonstrate that the 9-mer represents the naturally processed peptide from FGF-5. Antigenic activity of acid-stripped peptides from COS-A3 cells expressing FGF-5 was

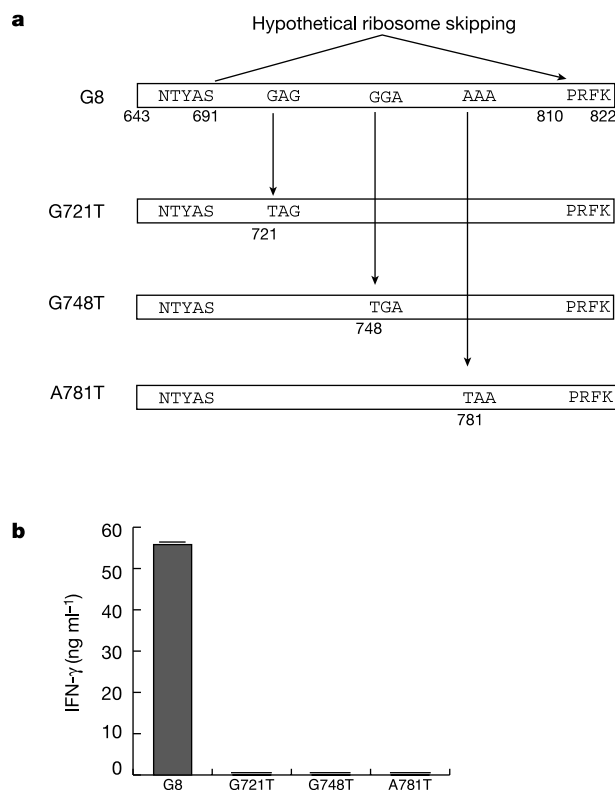


Figure 4 Determinant generation is not the result of ribosome skipping or RNA splicing. **a**, Termination codons were introduced by point mutagenesis into the intervening sequence of the G8 plasmid as indicated. **b**, Each stop codon prevented C2 activation by HLA-A3⁺ APCs transfected with these plasmids. Error bars show the standard deviation of duplicate IFN- γ determinations.

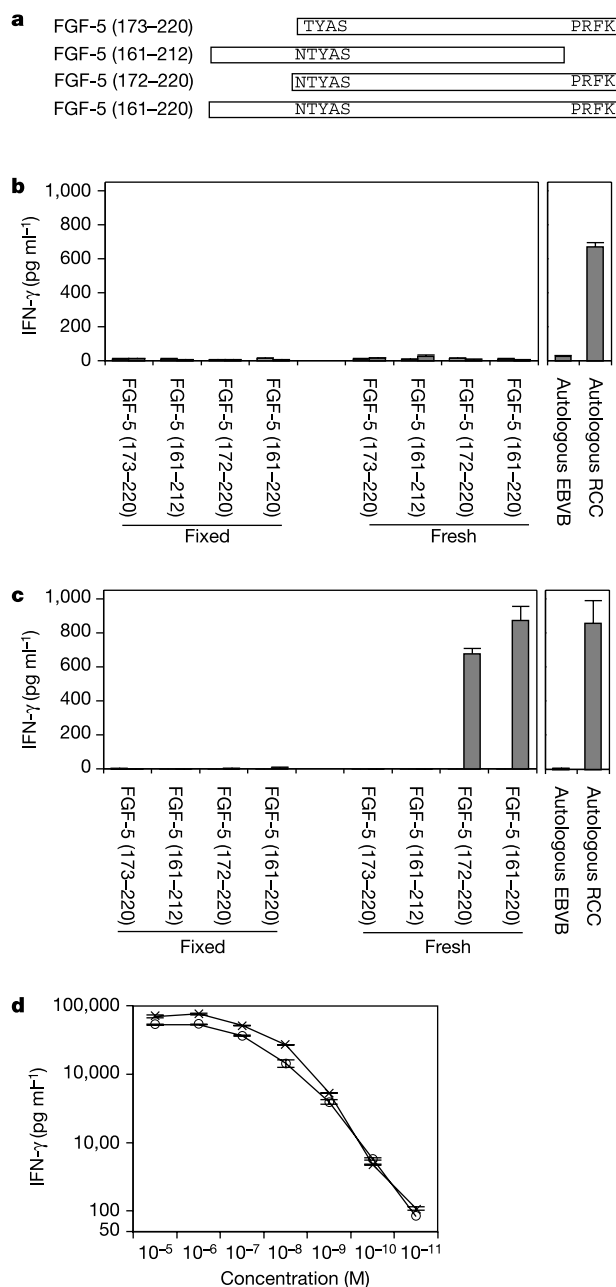


Figure 5 Presentation of synthetic peptides by EBV-B cells. **a**, Diagram of the extended FGF-5 synthetic peptides used in **b** and **c**. The peptides indicated were synthesized and tested for presentation by live or lightly fixed EBV-B cells. FGF-5 (173–220) was a 48-mer peptide lacking the first N-terminal asparagine of NTYAS but including C-terminal PRFK. FGF-5 (161–212) was a 52-mer peptide including NTYAS but lacking PRFK. FGF-5 (172–220) was a 49-mer peptide starting with NTYAS and ending with PRFK. FGF-5 (161–220) was a 60-mer peptide encoded by the G8 plasmid in Fig. 1a. **b**, **c**, These peptides were pulsed onto either fixed or fresh EBV-B cells that were either HLA-A3⁻ (open bars) or HLA-A3⁺ (filled bars) and were tested for recognition by control HLA-A3 restricted RCC-reactive CTLs from the same patient not recognizing FGF-5 (**b**) or by FGF-5-reactive C2 CTLs (**c**). The functional specificity of each CTL line is shown in the right-hand panel with the use of the autologous EBV-B cell line and the autologous RCC cell line as targets. **d**, Influence of fixing on HLA-A3. Adequate expression and function of fresh (crosses) and fixed (circles) EBV-B cells are shown by the similar capacity to present the 9-mer determinant NTYASPRFK. Error bars show the standard deviation of duplicate IFN- γ determinations. Similar results were obtained in two additional independent experiments.

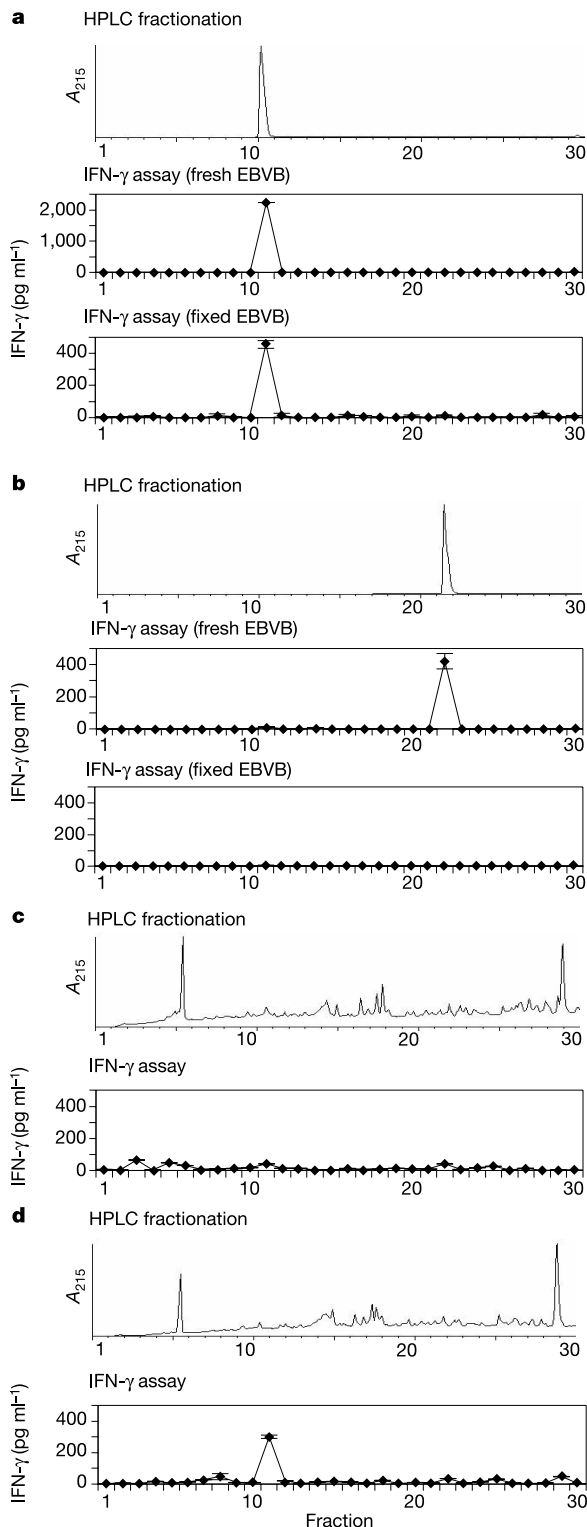


Figure 6 HPLC analysis of synthetic and acid-stripped peptides. **a**, Synthetic 9-mer (NTYASPRFK); **b**, synthetic 49-mer (172–220); **c**, acid-stripped peptides from COS-A3; **d**, acid-stripped peptides from COS-A3/FGF-5. Peptides were fractionated on a C18 HPLC column and the fractions were tested for recognition by C2. Similar results were reproduced by using a column and HPLC system that had never been used for the analysis of the 9-mer determinant, eliminating contamination by the 9-mer as a source of the antigenicity in **d**, fraction 11. Fresh EBV-B or fixed EBV-B cells were used as APCs in **a** and **b**, and fresh EBV-B cells were used as APCs in **c** and **d**. Error bars show the standard deviation of duplicate IFN- γ determinations.

exclusively recovered from fraction 11 (Fig. 6d). By contrast, none of the fractions from non-FGF5 expressing cells showed significant antigenic activity (Fig. 6c). These results indicate that the naturally processed FGF-5 determinant is the covalently linked 9-mer and not the 49-mer or the non-covalently bonded peptides NTYAS and PRFK, which should elute in fractions 28 and 5, respectively (not shown).

Given the highly unusual origins of the 9-mer determinant, it was of interest to characterize the antigen-processing pathway used to generate it from full-length FGF-5. To facilitate this analysis we transduced HLA-A2, gp100 and the MHC class II transactivator (CIITA) gene⁸ into the autologous RCC line to create a tumour target capable of presenting not only its autologous antigens to C2, but also class I and II-restricted determinants from gp100 to control T-cell clones that recognize this melanoma antigen (class II gp100 recognition is restricted by DR β 1*0401, naturally present on the RCC). Presentation of both MHC class I-restricted peptides was blocked by the proteasome inhibitor clasto-lactacystin β -lactone or by expressing ICP47 to inhibit peptide transport from cytosol to endoplasmic reticulum mediated by the transporter associated with antigen processing (TAP) (Supplementary Fig. 1). As a control, MHC class II-restricted gp100 recognition was unaffected.

On the basis of these data, we conclude that the NTYASPRFK peptide is generated by protein splicing from longer biosynthetic or synthetic precursors. Lower organisms are known to splice proteins by means of inteins, self-liberating proteins that ligate the flanking extein sequences in the absence of auxiliary factors⁴. This is unlikely to explain our findings. The smallest known intein consists of 134 amino acids⁹, and most are much longer. The intervening sequence in FGF-5 is 40 amino acids, and by gene truncation analysis it is possible to shorten it to as few as 18 amino acids while preserving splicing. Inteins typically have N-terminal Cys, Ser or Ala residues and C-terminal His or Asn residues and are inserted into host proteins next to Cys, Ser or Thr. None of these motifs is present in FGF-5. The splicing we observe is more likely to occur by reverse proteolysis, as described for concanavalin A^{3,10}. The dependence of antigen presentation on the proteasome and on TAP, and the successful splicing of plasmid-encoded truncated FGF-5-based polypeptides lacking leader sequences, suggest that splicing occurs in the cytosol. In the presence of the leader sequence on FGF-5, delivery of FGF-5 to the cytosol for splicing might require the endoplasmic-reticulum-associated degradation pathway¹¹, which is known to be used for antigen presentation^{12,13}. It remains to be established whether protein splicing in human cells occurs only during protein degradation or is used to generate mature stable substrates, as occurs with concanavalin A in plants.

Our findings show that the immune system monitors non-contiguous peptide sequences generated post-translationally. This capability represents an enormous increase in the ability of CTLs to recognize self and foreign proteins. Although this should enhance immune surveillance, from the practical standpoint it complicates the task of identifying peptide ligands recognized by tumour-specific and pathogen-specific CTLs. In a broader context, the existence of protein splicing in vertebrates greatly increases the cell's options for converting genetic information into proteins. □

Methods

CTL recognition assay

FGF-5 nucleotide modifications: deletion mutants and alanine substitution mutants were generated by polymerase chain reaction (PCR) from the wild-type FGF-5 sequence (GenBank accession no. AF535149), which was obtained from a renal cancer cell line derived from the patient who also provided C2 CTLs; they were cloned into the eukaryotic expression vector pME18S, sequenced and used for the assay as described previously².

Peptide pulsing: for Fig. 3, EBV-B cells were incubated for 30 min at 24 °C with candidate peptides at 10 μ M; for Fig. 5 the conditions were 3 h at 37 °C with peptides at 10 μ M. After the incubation, EBV-B cells were washed and CTLs were added. IFN- γ in the supernatant was measured 20–24 h later by ELISA. Peptide pulsing was done in RPMI

medium 1640 (RPMI) and the co-culture was done in RPMI with 10% fetal bovine serum (FBS). Where indicated, EBV-B cells were fixed in 1% formaldehyde in PBS for 10 min on ice, then washed in PBS. Peptides were synthesized either on the Pioneer Peptide Synthesizer (PE Biosystem) or on the AMS 222 multiple peptide synthesizer (Gibson) using standard fluoren-9-yl-methoxycarbonyl chemistry. The molecular masses of peptides were verified on a mass spectrometer (Bio-Synthesis Inc.).

Site-directed mutagenesis

Point mutations were introduced with a Quick Change Site-Directed Mutagenesis kit from Stratagene in accordance with the manufacturer's protocol.

HPLC fractionation of synthetic and acid-stripped peptides

COS-A3 and COS-A3/FGF-5 were prepared by retrovirally transducing COS-7 cells. Each cell line was cultured in T175 flasks and the peptides were stripped by adding to each flask 7.5 ml of citrate buffer (0.13 M citric acid, 0.056 M sodium phosphate dibasic pH 3.1) for 90 s. Peptide solutions (from about 4×10^9 cells, about 1,200 ml each) were prepared and were concentrated on a Sep-pak Plus C18 column (Waters). After freeze-drying and resolubilization into 200 μ l of a mixture of 5% acetonitrile and 0.05% (v/v) trifluoroacetic acid, peptides were fractionated on a C18 HPLC column (218TP54; Grace Vydak) using a linear gradient from 5% acetonitrile with 0.05% (v/v) TFA to 40% acetonitrile with 0.05% (v/v) TFA (1% min⁻¹ and 1 ml min⁻¹). Fractions (1 ml each) were collected, freeze-dried, reconstituted in 100 μ l RPMI and added to EBV-B cells. After incubation for 3 h, CTLs in 100 μ l RPMI with 20% FBS were added and IFN- γ was measured at 20 h. Synthetic peptides (20 μ g 9-mer and 100 μ g 49-mer) were fractionated by HPLC and assayed in the same way as above. All the 9-mer fractions were diluted 1:100 before the assay.

Antigen-processing pathway analysis

The RCC cell line was prepared by retroviral transduction with genes for HLA-A2, CIITA and a melanoma antigen gp100.

Treatment with clasto-lactacystin β -lactone (Calbiochem): RCC cells were treated with citric acid buffer (pH 3.1) for 90 s and washed, then cultured in 10 μ M clasto-lactacystin β -lactone for 3 h. After washing, T cells were added and IFN- γ was measured at 20 h.

TAP-1 dependence analysis: the same RCC line as above was infected with either an adenovirus encoding green fluorescent protein or the TAP-1 inhibitor ICP47 (multiplicity of infection = 100; 2 h). After washing, overnight culture and treatment of the RCC cells with the citric acid buffer as above, T cells were added. IFN- γ was measured at 20 h.

Received 11 September; accepted 10 November 2003; doi:10.1038/nature02240.

1. Yewdell, J. W., Reits, E. & Neefjes, J. Quantitating the MHC class I antigen processing pathway. *Nature Rev. Immunol.* **3**, 952–961 (2003).
2. Hanada, K., Perry-Lalley, D. M., Ohnmacht, G. A., Bettinotti, M. P. & Yang, J. C. Identification of fibroblast growth factor-5 as an overexpressed antigen in multiple human adenocarcinomas. *Cancer Res.* **61**, 5511–5516 (2001).
3. Carrington, D. M., Auffret, A. & Hanke, D. E. Polypeptide ligation occurs during post-translational modification of concanavalin A. *Nature* **313**, 64–67 (1985).
4. Paulus, H. Protein splicing and related forms of protein autoprocesing. *Annu. Rev. Biochem.* **69**, 447–496 (2000).
5. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. & Stevanovic, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50**, 213–219 (1999).
6. Huang, W. M. *et al.* A persistent untranslated sequence within bacteriophage T4 DNA topoisomerase gene 60. *Science* **239**, 1005–1012 (1988).
7. Burset, M., Seledtsov, I. A. & Soloviyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364–4375 (2000).
8. Steimle, V., Otten, L. A., Zufferey, M. & Mach, B. Complementation cloning of an MHC class II transactivator mutated in hereditary MHC class II deficiency (or bare lymphocyte syndrome). *Cell* **75**, 135–146 (1993).
9. Evans, T. C. Jr, Benner, J. & Xu, M. Q. The *in vitro* ligation of bacterially expressed proteins using an intein from *Methanobacterium thermoautotrophicum*. *J. Biol. Chem.* **274**, 3923–3926 (1999).
10. Min, W. & Jones, D. H. *In vitro* splicing of concanavalin A is catalyzed by asparaginyl endopeptidase. *Nature Struct. Biol.* **1**, 502–504 (1994).
11. McCracken, A. A. & Brodsky, J. L. Evolving questions and paradigm shifts in endoplasmic-reticulum-associated degradation (ERAD). *BioEssays* **25**, 868–877 (2003).
12. Bacik, I. *et al.* Introduction of a glycosylation site into a secreted protein provides evidence for an alternative antigen processing pathway: transport of precursors of major histocompatibility complex class I-restricted peptides from the endoplasmic reticulum to the cytosol. *J. Exp. Med.* **186**, 479–487 (1997).
13. Mosse, C. A. *et al.* The class I antigen-processing pathway for the membrane protein tyrosinase involves translation in the endoplasmic reticulum and processing in the cytosol. *J. Exp. Med.* **187**, 37–48 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. R. Bennink and P. F. Robbins for suggestions, comments and encouragement; J. P. Riley and M. R. Parkhurst for the peptide synthesis, and S. A. Rosenberg for continuous support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.H. (khanada@nih.gov) or J.C.Y. (jamesyang@mail.nih.gov).

Ras regulates assembly of mitogenic signalling complexes through the effector protein IMP

Sharon A. Matheny¹, Chiyuan Chen¹, Robert L. Kortum², Gina L. Razidlo², Robert E. Lewis² & Michael A. White¹

¹Department of Cell Biology, UT Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9039, USA

²Eppley Institute for Research in Cancer and Allied Diseases, Department of Pathology, University of Nebraska Medical Center, Omaha, Nebraska 68198-6805, USA

The signal transduction cascade comprising Raf, mitogen-activated protein (MAP) kinase kinase (MEK) and MAP kinase is a Ras effector pathway that mediates diverse cellular responses to environmental cues and contributes to Ras-dependent oncogenic transformation. Here we report that the Ras effector protein Impedes Mitogenic signal Propagation (IMP) modulates sensitivity of the MAP kinase cascade to stimulus-dependent activation by limiting functional assembly of the core enzymatic components through the inactivation of KSR, a scaffold/adaptor protein that couples activated Raf to its substrate MEK¹. IMP is a Ras-responsive E3 ubiquitin ligase that, on activation of Ras, is modified by auto-polyubiquitination, which releases the inhibition of Raf–MEK complex formation. Thus, Ras activates the MAP kinase cascade through simultaneous dual effector interactions: induction of Raf kinase activity and derepression of Raf–MEK complex formation. IMP depletion results in increased stimulus-dependent MEK activation without alterations in the timing or duration of the response. These observations suggest that IMP functions as a threshold modulator, controlling sensitivity of the cascade to stimulus and providing a mechanism to allow adaptive behaviour of the cascade in chronic or complex signalling environments.

IMP was identified as a Ras-interacting protein from a *Xenopus* yeast two-hybrid screen using the Ras effector loop variant G12V/E37G (compromised for Raf interaction) as bait (GenBank accession number AY331413). The association was specific, as IMP did not interact with other effector loop variants, nor with any other Ras family G protein tested (Fig. 1a and data not shown). Truncation analysis identified a domain of 104 amino acids as minimally sufficient to mediate the two-hybrid interaction with Ras. Using

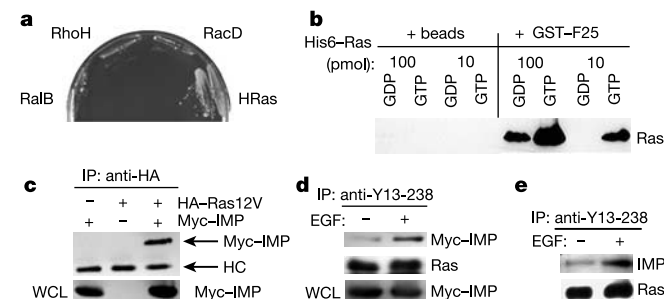


Figure 1 IMP is a Ras effector. **a**, IMP selectively associates with H-Ras in yeast. **b**, Recombinant His₆-Ras-GTP binds directly to GST-F25 (corresponding to residues 273–377 of IMP) *in vitro*. **c**, Co-immunoprecipitation of Myc-IMP with HA-Ras12V from NIH3T3 cells stably expressing HA-RasG12V. **d**, Stimulus-dependent co-immunoprecipitation of Myc-IMP with endogenous Ras from HeLa cells. **e**, Stimulus-dependent co-immunoprecipitation of endogenous IMP with endogenous Ras from HeLa cells.

recombinant proteins, we found that this domain binds directly to human H-Ras in a GTP-dependent manner (Fig. 1b). The full-length human orthologue (GenBank accession number AY332222), isolated from a Jurkat T-cell library, associated with oncogenic Ras and mitogen-activated Ras in cells (Fig. 1b–d). Notably, wild-type IMP formed a stimulus-induced complex with wild-type Ras in human cells (Fig. 1e), suggesting it is a Ras effector protein.

The primary structure of IMP suggests that it has a RING-H2 domain, followed by a ubiquitin-protease-like zinc-finger (UBP-ZnF) and leucine heptad repeats that are predicted to form a coiled-coil (SMART, Simple Modular Architecture Research Tool, <http://smart.embl-heidelberg.de>). This architecture is similar to the RING-B box coiled-coil (RBCC) family of proteins that includes the proto-oncogenes TIF-1 and PML². IMP was previously reported as a BRCA1-interacting protein 2 (Brp2) that binds the polybasic nuclear localization signal of BRCA1; however, the relevance of the interaction is unknown³. IMP is highly conserved across eukaryotes, with a single orthologue present in each species (Supplementary Fig. S1a). Multiple human tissue northern blot analysis

showed broad tissue expression (data not shown), as previously described in mouse³ and human.

Given that several Ras effector pathways collaborate with the Raf/MAP kinase cascade to induce oncogenic transformation⁴, we examined the potential contribution of IMP to Ras signalling. We expressed IMP together with the oncogenic Raf variant RafBxB⁵, or the constitutively active MEK variant MEKR4F, and examined the effect on downstream signal output. IMP inhibited focus formation of RafBxB-transformed NIH3T3 cells (data not shown), blocked Raf-induced activation of endogenous MEK and extracellular signal regulated kinase (ERK), and blocked Raf-induced accumulation of c-Fos; however, IMP did not inhibit MEKR4F signalling (Fig. 2a). The inhibitory effect of IMP seemed to be selective for Raf, as the activity of another MAP3K family protein kinase, MEKK3, was not affected by IMP expression (Fig. 2a). IMP blocked Raf signalling but did not seem to block activation of Raf, as neither a Ras-dependent activating and phosphorylation event on serine 338 (ref. 6 and Fig. 2b), nor Raf kinase activity (Fig. 2a, c), was inhibited by IMP. On the contrary, phosphorylation of S338 was enhanced. As IMP

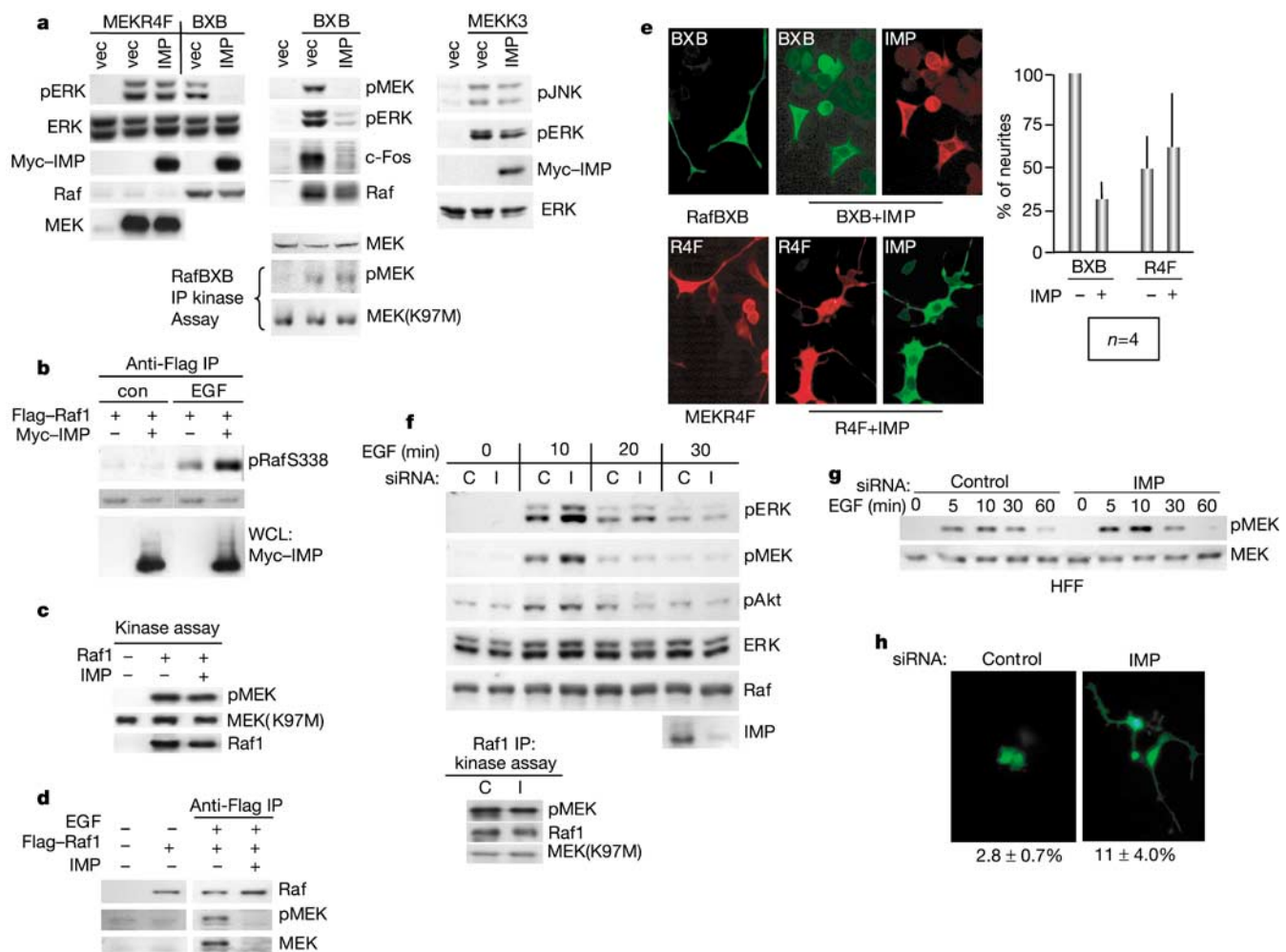


Figure 2 IMP impedes signal transmission from Raf to MEK. **a**, The effect of IMP expression on activation of MEK and ERK by RafBxB, activation of ERK by MEKR4F, and activation of ERK and JNK by MEKK3 was assessed by activation-specific (phosphorylation-specific) antibodies (pERK, pMEK and pJNK). The kinase activity of immunoprecipitated RafBxB was evaluated *in vitro* by using recombinant MEK(K97M) as a substrate. **b**, **c**, The effect of IMP expression on stimulus-dependent Raf1 phosphorylation and activation was evaluated with an antibody specific for Raf1 phosphorylation on S338 (**b**) and by *in vitro* kinase reactions (**c**), as described in **a**. WCL, whole-cell lysates. **d**, IMP blocks stimulus-dependent Raf–MEK association. **e**, IMP inhibits RafBxB- but not

MEKR4F-induced PC12 neurite-like extensions. Shown is the normalized percentage of cells that had extensions greater than two cell bodies in length. **f**, Depletion of IMP in HeLa cells. Cells were transfected with control (C) siRNA or siRNA targeting human IMP (I) and stimulated with EGF. Endogenous Raf1 kinase activity was assayed *in vitro* as described in **a**. **g**, Depletion of IMP in primary HFFs, treated as described in **f**. **h**, Depletion of IMP in PC12 cells. Cells were transfected with siRNAs plus GFP and treated with 10 ng ml^{−1} NGF. GFP-positive cells were scored for the presence of neurite-like extensions as described in **e**.

blocks ERK activation, this last observation was probably a consequence of uncoupling negative feedback inhibition of epidermal growth factor (EGF) signalling. However, IMP did block mitogen-induced association of Raf1 with endogenous MEK1 and MEK2

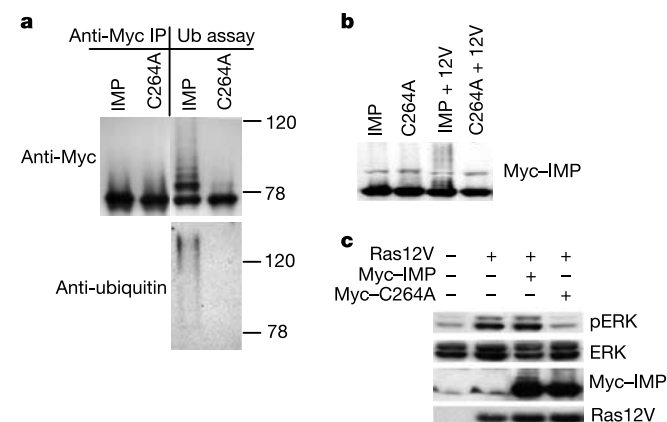


Figure 3 IMP is a Ras-sensitive RING-H2 E3 ubiquitin ligase. **a**, Immunoprecipitated Myc-IMP or Myc-IMP(C264A) was mixed with minimal essential purified ubiquitin ligase components, minus an E3 ligase, *in vitro*. IMP auto-ubiquitination was detected by immunoblotting. **b**, Ras12V enhances auto-ubiquitination of IMP. Whole-cell lysates were immunoblotted as shown. **c**, Ligase-defective IMP(C264A), but not wild-type IMP, can block ERK activation by oncogenic Ras.

(MEK1/2; Fig. 2d). Thus, IMP inhibits signal propagation through Raf by uncoupling activated Raf from its downstream substrates MEK1/2. This inhibition translates to downstream cellular responses, as shown by the capacity of IMP to block Raf-induced immediate early gene expression (Fig. 2a) and Raf-induced neurite-like differentiation of PC12 cells (Fig. 2e).

To validate the hypothesis that IMP is a negative regulator of Raf kinase signalling, we depleted IMP in various cell types. In human epithelial cells and primary human fibroblasts, depletion of IMP did not increase baseline ERK activity in serum-starved cells, but did increase the amplitude of the response of the MAP kinase cascade to mitogen stimulation (Fig. 2f, g). The duration of MEK activation was not appreciably altered, suggesting that IMP does not contribute to immediate early negative feedback control but, instead, limits the absolute capacity of the system to respond to ligand stimulation. EGF signalling was not generally increased on inhibition of IMP expression, as activation of AKT was unaffected (Fig. 2f). The capacity of IMP to modulate ERK activation negatively was also conserved in *Drosophila* S2 cells (Supplementary Fig. S2). These observations suggest that IMP adds impedance or resistance to signal propagation through Raf to MEK, thereby modulating the cellular threshold of sensitivity to stimulus. Consistent with this, we found that short interfering RNAs (siRNAs) targeting rat IMP enhanced the sensitivity of PC12 cells to suboptimal doses of nerve growth factor (NGF), as evaluated by morphological differentiation (Fig. 2h).

A potential functional motif in IMP is the putative RING-H2

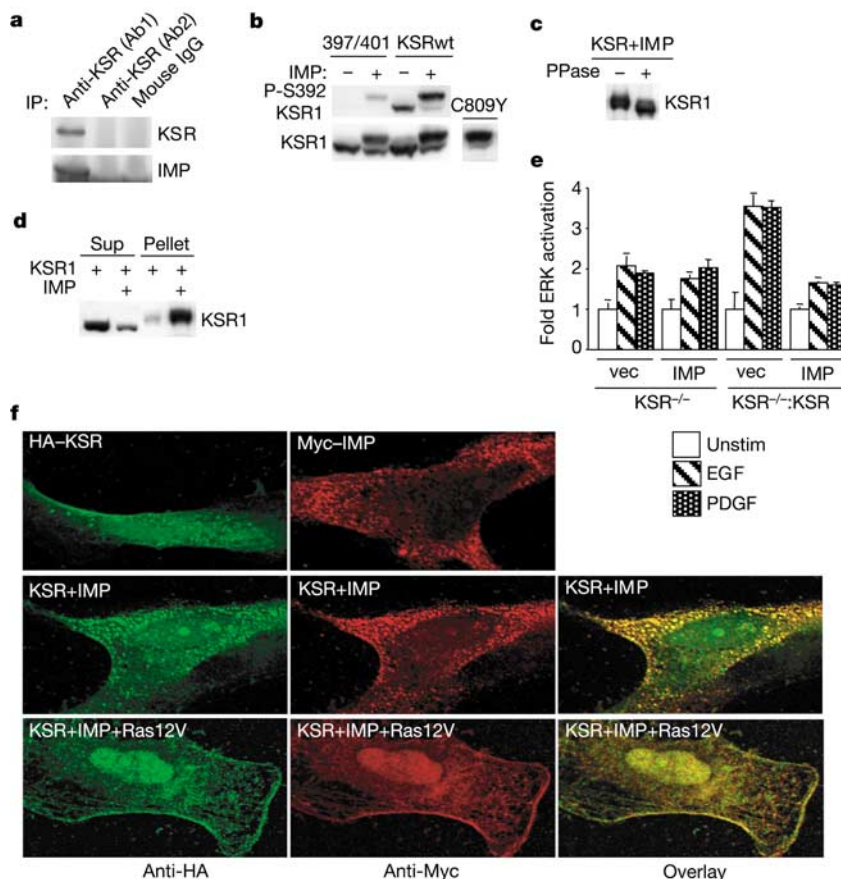


Figure 4 IMP inactivates KSR. **a**, Endogenous IMP co-immunoprecipitates with endogenous KSR from rat brain lysates. An ineffective KSR antibody (Ab2) and 'normal' IgG were used as controls. **b**, Effect of IMP expression on the mobility and inhibitory phosphorylation of KSR1. Whole-cell lysates were immunoblotted to detect KSR and the indicated KSR variants. Phosphorylation of S392 was detected with a phosphorylation-specific antibody. **c**, IMP induces hyperphosphorylation on KSR1. HA-KSR1 was

immunoprecipitated from the insoluble fraction (see **d**) and treated with phosphatase. **d**, High-molecular-mass KSR is insoluble in Triton. HEK293 cell lysates were separated into Triton-soluble and Triton-insoluble fractions. Equal amounts of protein were loaded. **e**, IMP inhibition of ERK activation is dependent on KSR. **f**, IMP and KSR1 colocalize in Ras-sensitive Triton-insoluble structures.

domain (Supplementary Fig. S1c). In many proteins, this domain facilitates protein ubiquitination through E3 ubiquitin ligase activity⁷. To examine whether IMP has E3 ubiquitin ligase activity, immunoprecipitated IMP or an IMP variant, in which the first cysteine in the RING-H2 consensus sequence was changed to alanine (C264A), was assayed for ligase activity by a standard *in vitro* reconstitution assay. IMP became polyubiquitinated *in vitro* in a RING-domain-dependent manner in the presence of recombinant E1, UbcH4 (E2) and ubiquitin (Fig. 3a). Thus, IMP is associated with E3 ubiquitin ligase activity and, like most E3 ligases, seems to be an auto-substrate *in vitro*.

We did not detect IMP-dependent modification of H-Ras, Raf1, MEK1 or ERK1 (data not shown). However, IMP is probably a *bona fide* target for its own E3 ligase activity in cells, as indicated by the appearance of a 'ladder' of high-molecular-mass IMP species in mitogen-stimulated cells and in cells expressing Ras12V together with IMP (Fig. 3b). This was not observed with the ligase-defective C264A mutant (Fig. 3b) and suggests that IMP E3 ligase activity is regulated by Ras. In contrast to RafBXB, IMP did not inhibit oncogenic Ras activation of MAP kinase. However, inactivation of the E3 ligase domain of IMP exposed inhibitory activity (Fig. 3c). These observations show that IMP inhibition of the Raf kinase cascade is Ras sensitive. Therefore, Ras is probably coupled to the MAP kinase cascade through dual effector interactions that mediate activation of Raf kinase together with derepression of Raf-MEK complex formation.

We considered that IMP might interfere with formation of the Raf-MEK complex through steric inhibition resulting from the competitive association of IMP with Raf and/or MEK, similar to the mechanism of action suggested for Raf kinase inhibitory protein (RKIP)⁸. However, we did not detect any substantial interaction of IMP with Raf1 or MEK1, either by yeast two-hybrid or by immunoprecipitation of ectopically expressed proteins (data not shown). KSR1 is an adaptor/scaffold protein that functions, at least in part, to couple Raf to MEK^{1,9,10}. As IMP uncouples Raf from MEK, we examined whether IMP affects the MAP kinase cascade at the level of KSR. In cells, IMP associated with KSR through the respective amino-terminal halves of the two proteins (Supplementary Fig. S3) and an endogenous KSR-IMP complex was detected (Fig. 4a). Although the dynamics of KSR function in cells is not well understood, studies suggest that KSR responds to signalling events. For example, KSR in cells is phosphorylated on several sites, some of which are modified in a mitogen-dependent manner^{9,11}. The serine/threonine protein kinase c-Tak-1 phosphorylates KSR on S392, and this modification can inactivate the scaffolding function of KSR in cells⁹. Furthermore, a KSR variant (C809Y), identified as the product of a loss-of-function KSR mutant in *Caenorhabditis elegans*¹², is hyperphosphorylated, fails to associate with MEK, and partitions to a Triton-insoluble fraction¹³.

We found that expression of IMP together with KSR resulted in a marked accumulation of a higher-molecular-mass species of KSR, mimicking the effects of the inactivating C809Y mutation (Fig. 4b). The IMP-induced modification of KSR was hyperphosphorylation, as indicated by phosphatase treatment (Fig. 4c), and, like KSR(C809Y), the hyperphosphorylated form partitioned to a Triton-insoluble cell fraction (Fig. 4d). Although phosphorylation of KSR on S392 was slightly enhanced by IMP, a KSR variant defective for c-Tak-1 association (I397A/V401A)⁹ showed that c-Tak-1-independent phosphorylation events on KSR accumulate in the presence of IMP (Fig. 4b). KSR^{-/-} fibroblasts isolated from knockout mice show reduced activation of ERK in response to various stimuli, as compared with wild-type cells¹, which can be enhanced by low ectopic expression of KSR (R.L.K and R.E.L., unpublished results). We used this system to examine whether IMP requires KSR to inhibit ERK activation. IMP did not inhibit EGF- or platelet-derived growth factor (PDGF)-stimulated ERK activation in KSR-null cells; however, IMP eliminated the capacity of KSR to

enhance ERK activation (Fig. 4e). Although the mechanistic consequences of the IMP-induced modifications of KSR are unknown, they are consistent with the behaviour of inactive KSR and suggest that IMP maintains KSR in an inactivated state. Immunostaining showed colocalization of IMP and KSR in Triton-resistant punctate structures (Fig. 4f), suggesting that IMP may recruit KSR to microdomains inaccessible to activators of KSR function. Expression of oncogenic Ras reversed this phenotype, further supporting the hypothesis that Ras inactivates IMP (Fig. 4f).

In summary, we have identified IMP, a Ras effector that negatively regulates MAP kinase activation by limiting the formation of Raf-MEK complexes. The mechanism of inhibition seems to be inactivation of the KSR1 adaptor/scaffold protein, showing that, in addition to promoting signal transmission, scaffold proteins can function to restrict signal propagation. Ras can inactivate IMP through induction of IMP auto-ubiquitination, facilitating KSR-dependent engagement of MEK by activated Raf. These observations show that Ras has dual effector inputs to the MAP kinase cascade: induction of Raf kinase activity, which is concomitant with derepression of KSR-dependent Raf-MEK complex formation. This relationship provides a mechanism to limit engagement of the MAP kinase cascade in the absence of Ras activation. MAP kinase activation contributes to many diverse cellular responses to the environment. The capacity to control the amplitude of this response through molecules such as IMP probably contributes to flexible and adaptive cellular behaviour in the context of complex regulatory signals. □

Methods

Expression vectors and antibodies

Expression vectors are given in the Supplementary Information. All primary antibodies were obtained from Santa Cruz Biotech except for those against HA.11 (Babco), Ras (Transduction Laboratories), Raf phosphorylated on S338 (Upstate Biotech), JNK1/2 phosphorylated on T183/Y185 (Cell Signalling), c-Fos (Upstate Biotech) and KSR (Transduction Laboratories). Antibodies against MEK1/2, MEK1/2 phosphorylated on S218/S222, ERK1/2 and ERK1/2 phosphorylated on T202/Y204 were from Sigma. All Raf immunodetection was done with a C-12 anti-Raf antibody (Santa Cruz Biotech) except for that in Fig. 2e, which used an anti-Raf1 monoclonal antibody (Pharmingen). Anti-IMP was generated against the carboxy-terminal peptide KLPSRKGRSKRGK (Biocarta). Antibody against KSR1 phosphorylated on S392 was generated against a peptide corresponding to S392 of KSR1 and its flanking residues (LRRTES(PO₃)/VPSD). We used the following secondary antibodies: horseradish peroxidase (HRP)-conjugated goat anti-mouse, HRP-conjugated goat anti-rabbit, fluorescein isothiocyanate (FITC)-conjugated goat anti-mouse and rhodamine-conjugated goat-rabbit (all from Jackson Laboratories). Anti-HA, anti-Y13-238 and anti-Myc9E10 agarose conjugates were from Santa Cruz Biotech. Anti-Flag M2 agarose was from Sigma.

Cell culture and transfection

Cell culture conditions are given in the Supplementary Information. Plasmid DNA was transfected by using calcium phosphate precipitation for NIH3T3 and HEK293 cells, and by using Lipofectamine 2000 (Promega) for HeLa and PC12 cells. siRNAs were transfected by using Oligofectamine (Promega). We treated S2 cells with double-stranded RNA as described¹⁴.

In vitro binding and ubiquitin ligase assays

The *in vitro* binding assay is described in the Supplementary Information. For the ubiquitin ligase assay, immunoprecipitated Myc-IMP or Myc-IMP(C264A) was added to purified assay components, comprising 5 μ M bovine ubiquitin (Sigma), 0.33 μ M E1, 13 nM recombinant human UbcH4, energy regeneration mix and ligase buffer (50 mM Tris (pH 7.5), 150 mM KCl and 1 mM MgCl₂), and mixed for 1.5 h at 37 °C.

RNA interference

HeLa cells and human foreskin fibroblasts (HFFs) were transfected with 100 pmol of annealed siRNA oligonucleotides targeting human IMP, or mouse Caveolin1 as a control, essentially as described¹⁵. To obtain high (>90%) transfection efficiencies with the primary fibroblasts, cells were briefly exposed to trypsin (Gibco), before transfection using Oligofectamine (Promega) as described¹⁶. After 24 h, the cells were stimulated with 1 ng ml⁻¹ EGF (Sigma) and then lysed by boiling in SDS-Tris for immunoblotting. Similar results were obtained with siRNAs targeting two non-overlapping sites on IMP. PC12 cells were transfected with 100 pmol of annealed siRNAs targeting mouse Caveolin1 or rat IMP, together with an expression plasmid encoding green fluorescent protein (GFP). After 48 h, cells were treated with 10 ng ml⁻¹ NGF (Sigma) and incubated for an additional 24 h. We scored GFP-positive cells for the presence of neurite-like extensions (Supplementary Information).

Immunoprecipitation

See Supplementary Information for the detailed protocol. In brief, cells were lysed in Nonidet P-40 (NP-40) buffer and cleared by centrifugation at 4 °C. Antibody-conjugated agarose was added to the supernatant and incubated at 4 °C for 3 h or overnight. Beads were washed four times in NP-40 buffer plus 500 mM NaCl. Co-immunoprecipitations were treated in the same way, except that the beads were washed in NP-40 buffer without salt.

Immunofluorescence

Cells were washed with PBS and fixed in 3.7% formaldehyde. Permeabilization, washes and antibody dilutions were done with 1% calf serum and 0.25% Triton X-100 in PBS. Primary antibodies were diluted 1:80. FITC-conjugated goat anti-mouse was used at a dilution of 1:300, and goat anti-rabbit was used at 1:1,000. For the experiments in Fig. 4f, the cells were washed with PBS then incubated in cold 0.1% Triton in PBS for 10 min at 4 °C. The cells were then fixed and immunostained as described above. All images were captured at a magnification of $\times 40$ on a confocal microscope (Leica).

For fluorescence *in situ* ERK activation assays (Fig. 4e), cells were stained with antibodies against phosphorylated ERK1/2 (Cell Signalling, diluted 1:100) and antibodies against ERK1 (Santa Cruz, 1:100), followed by AlexaFluor-680-conjugated anti-mouse (Molecular Probes, 1:100) and IRDye800-conjugated anti-rabbit (Rockland, 1:100) secondary antibodies. Signal detection and quantification were done by the Odyssey system (Li-Cor) as directed by the manufacturer.

In situ ERK activation assay

We infected KSR1^{-/-} cells with bicistronic retroviruses encoding KSR1-IRES-GFP and IMP-IRES-YFP or control viruses, and sorted them for green and yellow fluorescence by FACS. Sorted cells were seeded at 1.5×10^4 cells per well in 96-well plates 24 h before analysis. Cells at 70% confluence were deprived of serum for 4 h, and treated with 100 ng ml⁻¹ EGF or 25 ng ml⁻¹ PDGF for 5 min, before being assayed for ERK activation by a quantitative fluorescence *in situ* plate assay.

Phosphatase and Raf kinase assays

See Supplementary Information for the phosphatase assays. For the Raf kinase assays, Raf1 or RafBXB was immunoprecipitated and washed five times in NP-40 buffer, once in PBS, and once in 25 mM HEPES (pH 7.5) and 10 mM MgCl₂. The beads were then added to kinase assay reagents comprising 10 mM MgCl₂, 83 μ M ATP, 25 mM HEPES (pH 7.5) and 50 ng μ l⁻¹ recombinant MEK (K94A). The reaction was incubated at 30 °C for 40 min with regular mixing, and stopped on the addition of 2 $\times$ sample buffer.

Received 1 October; accepted 14 November 2003; doi:10.1038/nature02237.

1. Nguyen, A. *et al.* Kinase suppressor of Ras (KSR) is a scaffold which facilitates mitogen-activated protein kinase activation *in vivo*. *Mol. Cell. Biol.* **22**, 3035–3045 (2002).
2. Jensen, K., Shiels, C. & Freemont, P. S. PML protein isoforms and the RBCC/TRIM motif. *Oncogene* **20**, 7223–7233 (2001).
3. Li, S. *et al.* Identification of a novel cytoplasmic protein that specifically binds to nuclear localization signal motifs. *J. Biol. Chem.* **273**, 6183–6189 (1998).
4. White, M. A. *et al.* Multiple Ras functions can contribute to mammalian cell transformation. *Cell* **80**, 533–541 (1995).
5. Bruder, J. T., Heidecker, G. & Rapp, U. R. Serum-, TPA-, and Ras-induced expression from Ap-1/Ets-driven promoters requires Raf-1 kinase. *Genes Dev.* **6**, 545–556 (1992).
6. Diaz, B. *et al.* Phosphorylation of Raf-1 serine 338/serine 339 is an essential regulatory event for Ras-dependent activation and biological signalling. *Mol. Cell. Biol.* **17**, 4509–4516 (1997).
7. Joazeiro, C. A. & Weissman, A. M. RING finger proteins: mediators of ubiquitin ligase activity. *Cell* **102**, 549–552 (2000).
8. Yeung, K. *et al.* Suppression of Raf-1 kinase activity and MAP kinase signalling by RKIP. *Nature* **401**, 173–177 (1999).
9. Muller, J. S. O., Copeland, T., Piwnicka-Worms, H. & Morrison, D. K. C-TAK1 regulates Ras signalling by phosphorylating the MAPK scaffold, KSR1. *Mol. Cell* **8**, 983–993 (2001).
10. Roy, F., Laberge, G., Douziech, M., Ferland-McCollough, D. & Therrien, M. KSR is a scaffold required for activation of the ERK/MAPK module. *Genes Dev.* **16**, 427–438 (2002).
11. Brennan, J. A., Volle, D. J., Chaika, O. V. & Lewis, R. E. Phosphorylation regulates the nucleocytoplasmic distribution of kinase suppressor of Ras. *J. Biol. Chem.* **277**, 5369–5377 (2002).
12. Sundaram, M. & Han, M. The *C. elegans* KSR-1 gene encodes a novel Raf-related kinase involved in Ras-mediated signal transduction. *Cell* **83**, 889–901 (1995).
13. Stewart, S. *et al.* Kinase suppressor of Ras forms a multiprotein signalling complex and modulates MEK localization. *Mol. Cell. Biol.* **19**, 5523–5534 (1999).
14. Clemens, J. C. *et al.* Use of double-stranded RNA interference in *Drosophila* cell lines to dissect signal transduction pathways. *Proc. Natl Acad. Sci. USA* **97**, 6499–6503 (2000).
15. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
16. Chien, Y. & White, M. A. RAL GTPases are linchpin modulators of human tumour-cell proliferation and survival. *EMBO Rep.* **4**, 800–806 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by grants from the National Cancer Institute (to M.A.W. and to R.E.L.) and the Welch Foundation, and by an Idea Development award (to M.A.W.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.A.W. (michael.white@utsouthwestern.edu).

Mustard oils and cannabinoids excite sensory nerve fibres through the TRP channel ANKTM1

Sven-Eric Jordt¹, Diana M. Bautista¹, Huai-hu Chuang¹, David D. McKemy¹, Peter M. Zygmunt³, Edward D. Högestätt³, Ian D. Meng^{2*} & David Julius¹

¹Department of Cellular and Molecular Pharmacology and ²Department of Neurology, University of California, San Francisco, California 94143-2140, USA

³Department of Clinical Pharmacology, Institute of Laboratory Medicine, Lund University Hospital, SE-221 85 Lund, Sweden

* Present address: Department of Physiology, College of Osteopathic Medicine, University of New England, 11 Hills Beach Road, Biddeford, Maine 04005, USA

Wasabi, horseradish and mustard owe their pungency to isothiocyanate compounds. Topical application of mustard oil (allyl isothiocyanate) to the skin activates underlying sensory nerve endings, thereby producing pain, inflammation and robust hypersensitivity to thermal and mechanical stimuli^{1,2}. Despite their widespread use in both the kitchen and the laboratory, the molecular mechanism through which isothiocyanates mediate their effects remains unknown. Here we show that mustard oil depolarizes a subpopulation of primary sensory neurons that are also activated by capsaicin, the pungent ingredient in chilli peppers, and by Δ^9 -tetrahydrocannabinol (THC), the psychoactive component of marijuana. Both allyl isothiocyanate and THC mediate their excitatory effects by activating ANKTM1, a member of the transient receptor potential (TRP) ion channel family recently implicated in the detection of noxious cold^{3,4}. These findings identify a cellular and molecular target for the pungent action of mustard oils and support an emerging role for TRP channels as ionotropic cannabinoid receptors^{5–8}.

Capsaicin and mustard oil are plant-derived natural products that elicit pain and inflammation when applied to the skin^{1,2}. These effects are produced by the depolarization of sensory nerve fibres that detect noxious stimuli (so-called nociceptors). In addition to transmitting 'pain signals' to the spinal cord, nociceptors may also release peptides—such as substance P and calcitonin-gene-related peptide (CGRP)—peripherally to produce vascular leakage and vasodilation, leading to inflammation and tenderness at the site of irritant application^{9,10}. Capsaicin mediates these effects by activating the vanilloid receptor TRPV1, a heat-sensitive cation channel on nociceptor terminals^{11–13}. TRPV1 belongs to the greater family of TRP channels, many of which are key components of signal transduction pathways in a variety of sensory systems^{14–16}.

Topical administration of mustard oil also produces nociceptor excitation, but whether this occurs through a direct action on sensory neurons or through a specific membrane receptor is not known¹⁷. TRPV1-deficient mice retain sensitivity to mustard oil, indicating that capsaicin and isothiocyanates excite nociceptors through distinct molecular mechanisms¹⁸. However, inflammatory responses to these irritants show partial cross-desensitization, suggesting that they act through convergent cellular signalling pathways^{19,20}.

To identify the cellular site of action of mustard oil, we used calcium imaging to ask whether allyl isothiocyanate (20 μ M) excites dissociated sensory neurons from rat trigeminal ganglia. Approximately 35% of cultured neurons showed rapid and robust increases in intracellular free calcium with the application of this compound (Fig. 1). Subsequent exposure to capsaicin (1 μ M) excited a larger cohort of neurons ($\sim 55\%$), encompassing all of the mustard-oil-responsive cells. Capsaicin-evoked responses result from an influx of divalent cations directly through the TRPV1 channel²¹. Similarly,

mustard-oil-evoked responses were eliminated when calcium was removed from the extracellular medium (not shown), consistent with a mechanism involving an influx of calcium across the plasma membrane.

The plant-derived cannabinoid receptor agonists THC and cannabidiol have recently been shown to relax hepatic or mesenteric arteries *in vitro* by activating capsaicin-sensitive, CGRP-containing perivascular sensory nerve endings that innervate the smooth muscle²². This effect is not inhibited by antagonists of known G-protein-coupled cannabinoid receptors, but is blocked by ruthenium red, an inhibitor of TRPV1 and certain other members of the TRP channel family. Moreover, THC-evoked vasorelaxation is dependent on extracellular calcium and persists in TRPV1-deficient mice²². Like THC, isothiocyanates induce endothelium-independent vasorelaxation²³, and activation of sensory nerve endings by these compounds is blocked by ruthenium red²⁴ and persists in TRPV1^{-/-} mice¹⁸. We therefore asked whether THC and allyl isothiocyanate excite the same population of neurons. Indeed, all THC-excitable cells responded to mustard oil (Fig. 1a, b), and responses to either agonist were blocked by ruthenium red (10 μ M) (Fig. 1c). A significant percentage ($\sim 35\%$) of mustard-oil-sensitive neurons were not excited by THC, but most of these cells showed relatively weak responses to mustard oil (25% less than THC-

sensitive cells, $P < 0.001$) and their failure to respond to THC may reflect the less robust effect of cannabinoids as excitatory agents for these cells (Fig. 1). Whole-cell patch-clamp recordings from trigeminal neurons confirmed our observation that a subpopulation of capsaicin-sensitive cells is dually responsive to mustard oil and THC (Fig. 1e). These agonists produced currents that reversed near 0 mV (consistent with the opening of non-selective cation channels), showed slight outward rectification and were blocked by ruthenium red (Fig. 1e).

Taken together, these results suggest that THC and mustard oil excite nociceptors through a similar, if not identical mechanism involving activation of a calcium-permeable, ruthenium-red-blockable channel on capsaicin-sensitive, CGRP-containing sensory neurons. In considering a potential molecular target for these agents, we realized that all of these characteristics are met by ANKTM1, a member of the TRP channel family recently proposed to function as a detector of noxious cold by peptidergic, TRPV1-expressing primary afferent neurons⁴. To test this possibility directly, we measured responses of rat ANKTM1-expressing human embryonic kidney (HEK) 293 cells to mustard oil. Indeed, application of mustard oil (20 μ M) produced large increases in intracellular free calcium, whereas capsaicin (1 μ M) had no effect (Fig. 2a). This response was specific to ANKTM1, as TRPV1- or

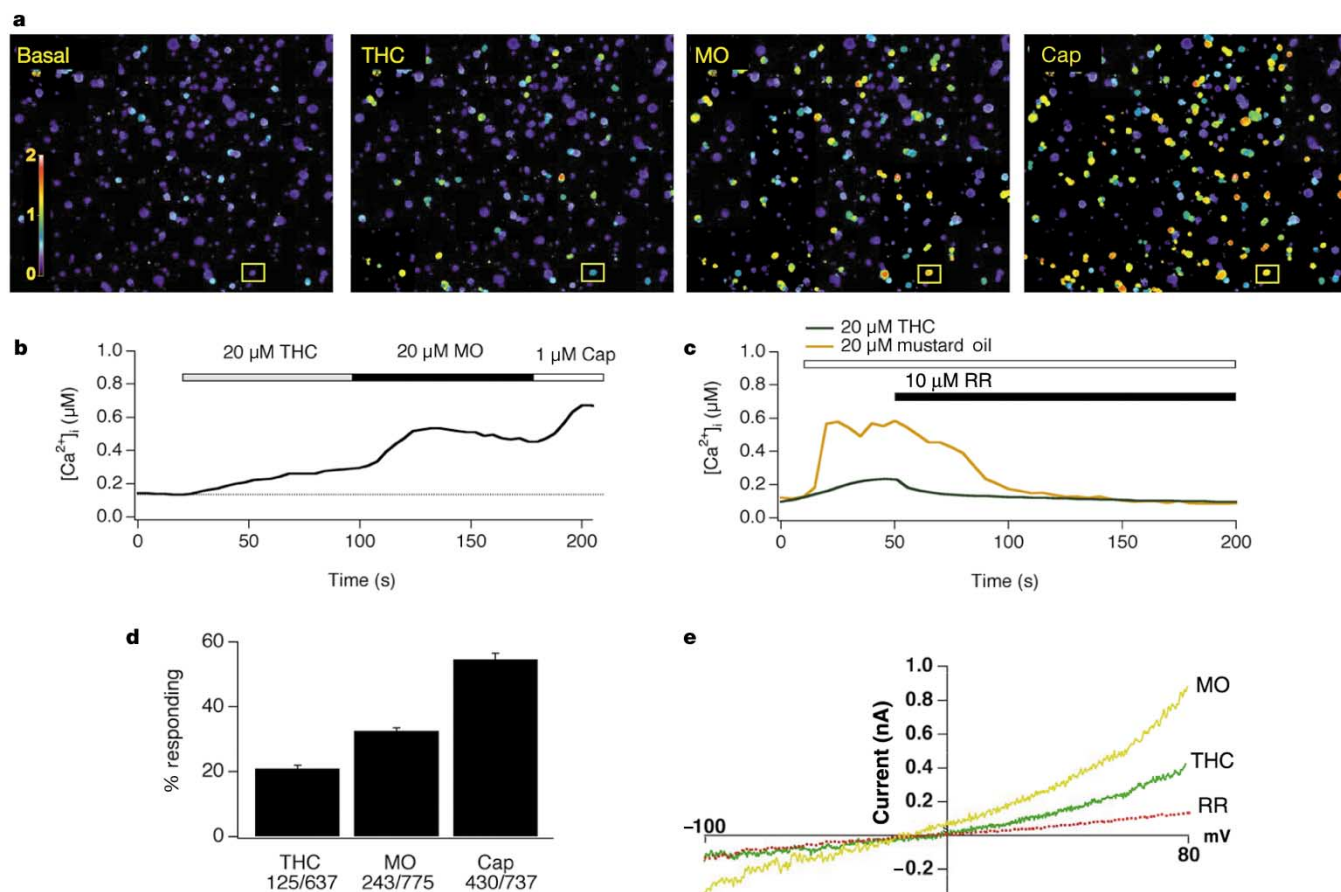


Figure 1 THC and mustard oil activate a subset of sensory neurons. **a**, Responses of dissociated rat trigeminal neurons to 20 μ M THC, 20 μ M mustard oil (MO) or 1 μ M capsaicin (Cap), as measured by Fura-2 ratiometric imaging (scale bar indicates the intracellular calcium concentration, $[Ca^{2+}]_i$, in μ M). **b**, $[Ca^{2+}]_i$ response in a representative neuron to THC, mustard oil and capsaicin (see box in part **a**). **c**, Calcium responses evoked by THC (green) and mustard oil (yellow) are blocked by 10 μ M ruthenium red (RR). As shown in panel **b**, responses are sustained in the absence of ruthenium red. Calcium increases returned to basal levels when agonists were washed

out (not shown). **d**, Percentage of cultured neurons exhibiting a stimulus-evoked rise in $[Ca^{2+}]_i$ in response to 20 μ M THC, 20 μ M mustard oil or 1 μ M capsaicin. **e**, Electrophysiological response of a representative neuron to 20 μ M THC (green) or 20 μ M mustard oil in the absence (yellow) or presence (red) of 5 μ M ruthenium red. Voltage ramps from -100 to $+80$ mV (200 ms) were applied every 2 s. THC-evoked currents display slight outward rectification. Mustard oil activates identical currents of larger amplitude in the same cell. Ruthenium red (10 μ M) blocks THC- and mustard-oil-evoked currents.

vector-transfected cells were unresponsive. Mustard oil also elicited large and reversible membrane currents in oocytes expressing ANKTM1, where dose-response analysis revealed a half-maximal effective concentration (EC_{50}) of $11 \pm 1 \mu\text{M}$ (Fig. 2b). To further verify the specificity of these ligands for ANKTM1, we tested a number of other TRP channels expressed by sensory neurons, including TRPV2, TRPV3 and TRPM8, none of which responded to mustard oil.

A variety of isothiocyanate compounds, including allyl, benzyl, phenylethyl, isopropyl and methyl isothiocyanate, constitute the main pungent ingredients in wasabi, yellow mustard, Brussels sprouts, nasturtium seeds and capers, respectively. Calcium imaging showed that each of these compounds was capable of activating the human ANKTM1 channel (Fig. 2c). Electrophysiological responses were also observed in oocytes expressing human ANKTM1, where large membrane currents were elicited in response to crude extracts prepared from brown mustard seed or wasabi paste (Fig. 2d), or to the isothiocyanates described above (Fig. 2e).

In view of our observation with trigeminal neurons, we asked whether cells heterologously expressing ANKTM1 also respond to THC or cannabinoil. Calcium imaging showed that both the rat and the human receptor could be activated by these ligands (Fig. 3a, b), consistent with the idea that isothiocyanates and cannabinoils excite sensory neurons through the same molecular mechanism. Recordings from ANKTM1-expressing oocytes showed that cannabinoils elicit membrane currents that have similar, if not identical, properties to those produced by mustard oil (Fig. 3a, c), including slight outward rectification and block by ruthenium red (not

shown). The synthetic cannabinoid receptor agonists HU-210 and CP55,940 do not produce neurogenic vasodilation¹⁷ and failed to activate ANKTM1 (not shown). Finally, responses to cannabinoils were significantly smaller than those elicited by equal concentrations of isothiocyanates (Fig. 3c). Taken together, these characteristics closely resemble those observed in sensory neurons or sensory nerve fibres.

In situ hybridization studies indicate that only 4% of cells within mouse dorsal root ganglia express ANKTM1 transcripts⁴, whereas we found that THC or mustard oil excites a larger population of neurons from rat trigeminal ganglia. This difference could reflect greater channel expression in rat trigeminal versus mouse dorsal root ganglia, upregulation of channel expression in culture and/or greater sensitivity of calcium imaging versus histological analysis. A significant part of this difference can be explained by the first of these possibilities, as *in situ* hybridization analysis that we carried out with trigeminal ganglia from newborn rats detected ANKTM1 expression in $20 \pm 2\%$ of neurons (644 of 3,221 cells) with a mean soma diameter of $18 \pm 1 \mu\text{m}$ ($n = 88$). These numbers are consistent with our functional studies showing that mustard oil and THC excite $\geq 20\%$ of cultured trigeminal neurons, encompassing 30–50% of small-diameter, capsaicin-sensitive cells. We therefore propose that mustard oils and THC mediate their excitatory effect on nociceptors through ANKTM1. Whether these compounds also act at other sites will be determined by the analysis of ANKTM1-deficient mice.

TRPV1 and TRPM8/CMR1 are thermosensitive channels that enable primary afferent neurons to detect heat and cold, respect-

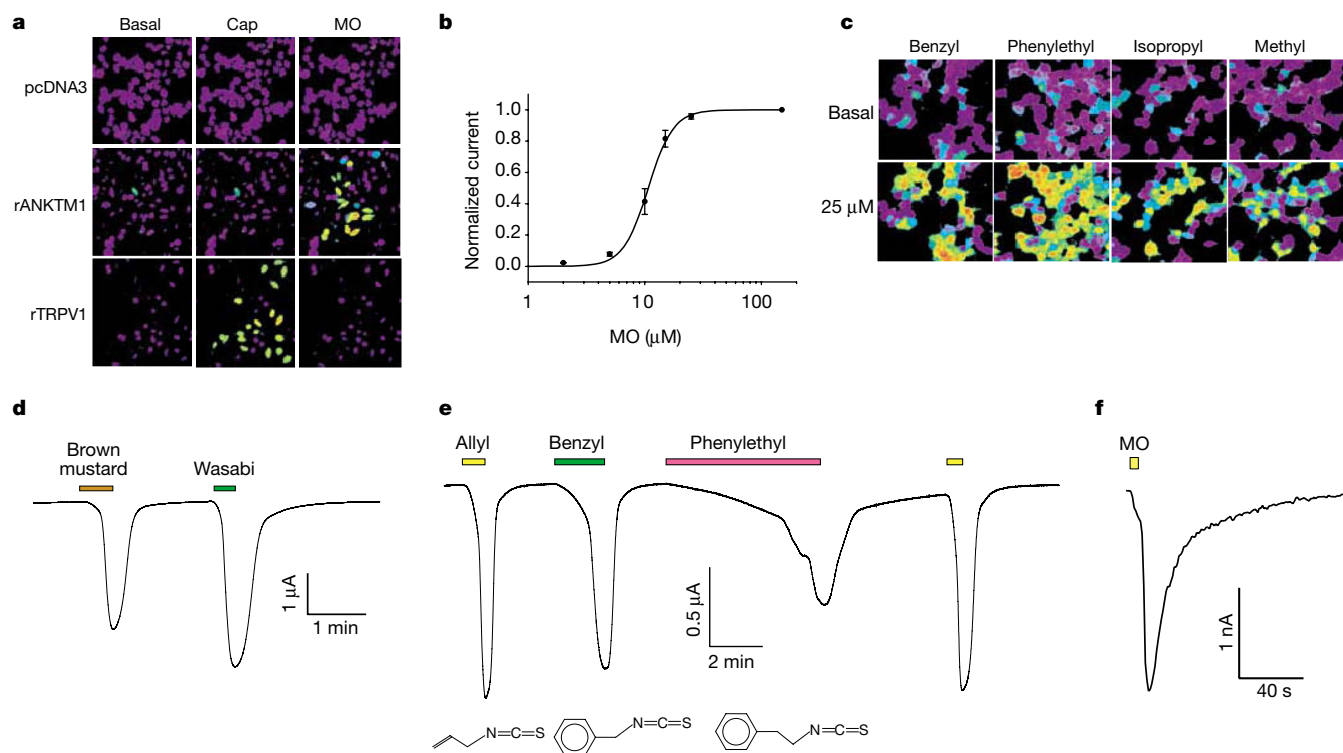


Figure 2 Activation of ANKTM1 by mustard oils in transfected mammalian cells and *Xenopus* oocytes. **a**, Mustard oil (allyl isothiocyanate, $20 \mu\text{M}$) activates calcium influx into rat ANKTM1-expressing HEK293 cells (middle row), measured by Fura-2 fluorescence. Cells transfected with vector (pcDNA3, top) or TRPV1 (bottom) are insensitive to mustard oil. Only TRPV1-expressing cells are activated by capsaicin ($1 \mu\text{M}$). **b**, Dose-response curve for activation of rat ANKTM1 by mustard oil in oocytes (holding potential, $V_h = -40 \text{ mV}$) recorded in nominally calcium-free ND96. Currents were normalized to maximal response evoked by $125 \mu\text{M}$ mustard oil. Half-maximal activation occurred at $11 \pm 1 \mu\text{M}$ ($n = 5$). **c**, Activation of human ANKTM1 by different pungent mustard oils

(benzyl-, phenylethyl-, isopropyl- and methyl isothiocyanate) in transfected HEK293 cells, measured by Fura-2 fluorescence. **d**, Activation of human ANKTM1 currents in oocytes by natural extracts of brown mustard and wasabi ($V_h = -60 \text{ mV}$). **e**, Representative trace from an oocyte expressing human ANKTM1, showing activation by allyl-, benzyl- or phenylethyl isothiocyanate ($10 \mu\text{M}$ each; $V_h = -60 \text{ mV}$). Agonist structures are indicated below the current trace. **f**, Activation of human ANKTM1 by mustard oil in transfected HEK293 cells, as measured by whole-cell patch clamp ($V_h = -60 \text{ mV}$). Brief application of mustard oil ($50 \mu\text{M}$ for 3 s) activated large inward currents with a mean amplitude of $2.3 \pm 0.3 \text{ nA}$ ($n = 11$).

ively^{12,25,26}. Activation of these receptors by plant-derived chemical agonists (capsaicin or menthol) produces sensations akin to those associated with the cognate thermal experiences. Recent studies suggest that ANKTM1 can be activated by thermal stimuli below 20 °C and may therefore serve as a detector of noxious cold⁴. Isothiocyanates produce a noxious sensation, as would be predicted for activation of a nociceptor, but the experience is qualitatively different from what one would expect if ANKTM1 simply served as a cold sensor. There are several possible explanations for this conundrum: (1) chemical and thermal agonists gate the channel and depolarize nociceptors differently to encode distinct psychophysical sensations; (2) isothiocyanates target several receptors, in addition to ANKTM1, to mediate their psychophysical effects; (3) activation of ANKTM1 alone produces a burning-like irritation, whereas the sensation of noxious cold requires the activation of additional channels, such as TRPM8; (4) a primary function of ANKTM1 may be to respond to chemical stimuli that produce pain and oedema, such as inflammatory peptides and neurotransmitters²⁷.

To address these possibilities, we carried out calcium imaging experiments with cultured rat trigeminal neurons to look for overlap between mustard oil and cold sensitivity. Among mustard-oil-sensitive neurons, 96% did not respond to a cold stimulus (5 °C). The 4% that were cold sensitive also responded to menthol, suggesting that activation of TRPM8 can account for cold sensitivity in these cells. When analysing neurons responding to cold, 95% responded to menthol and the remaining 5% were insensitive to both menthol and mustard oil. These results indicate that ANKTM1 is unlikely to underlie cold sensitivity in cultured trigeminal neurons, even in the small subpopulation of menthol-insensitive cells (Supplementary Fig. 1). We also examined the cold sensitivity of ANKTM1-expressing HEK293 cells and found no responses to 5 °C, even though they exhibited robust sensitivity to mustard oil. By contrast, TRPM8-expressing cells showed robust and reproducible responses to cold or menthol (Supplementary Fig. 1). In

view of our inability to observe reliable responses of ANKTM1-expressing cells to cold, we asked whether the effects of mustard oil might be augmented by cold. Instead we found that cooling of the perfusate reduced mustard-oil-evoked currents in ANKTM1-expressing oocytes (Supplementary Fig. 2), as one might expect for effects of cold on a non-thermally responsive ion channel.

Many TRP channels are activated or modulated downstream of neurotransmitter or growth factor receptors that stimulate phospholipase C (PLC) pathways^{14,16}. In considering alternative physiological roles for ANKTM1, we asked whether it might function as a receptor-operated channel in transfected HEK293 cells co-expressing the PLC-coupled M1 muscarinic acetylcholine receptor (mAChR). In cells expressing mAChR alone, activation by carbachol (100 μ M) produced a modest rise in intracellular calcium due to release from intracellular stores (Fig. 4a). However, in cells co-expressing ANKTM1, carbachol application produced a much more sustained and robust increase in cytoplasmic calcium that could be blocked by ruthenium red (Fig. 4a). All cells showing this enhanced calcium response (40/100) were also activated by mustard oil (20 μ M), whereas cells showing only the small release transient (60/100) were insensitive to mustard oil. Taken together, these results show that the large, sustained calcium increases were mediated through activation of ANKTM1. This was further supported by whole-cell patch-clamp recordings in which the application of acetylcholine (50 μ M) produced rapidly developing inward currents in cells co-transfected with ANKTM1 and mAChR, but not in cells expressing mAChR alone (Fig. 4b).

Recent studies suggest that TRP channels can be regulated by one or more consequences of PLC activation, including phosphatidylinositol biphosphate hydrolysis, the production of lipid metabolites or the release of calcium from intracellular stores^{14,16}. To ask whether calcium release is sufficient to activate ANKTM1, we treated cells with thapsigargin to increase intracellular calcium in a receptor-independent manner. In vector-transfected control cells, thapsigargin (1 μ M) induced a small and gradual increase in

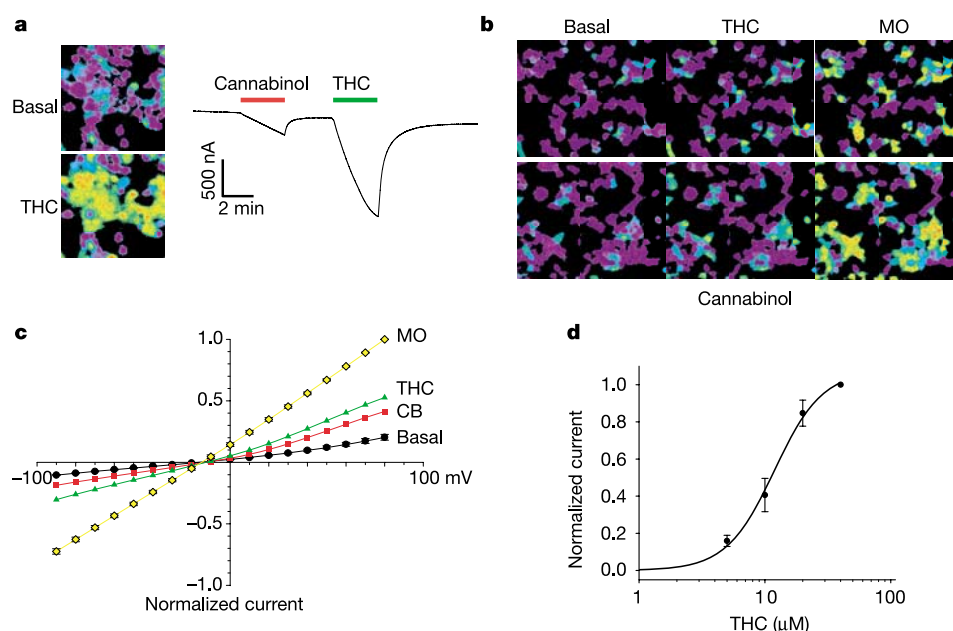


Figure 3 Activation of ANKTM1 by cannabinoids. **a**, Left, THC-induced calcium influx into HEK293 cells transfected with rat ANKTM1, as measured by Fura-2 fluorescence. Right, activation of inward currents at -60 mV by cannabinol (20 μ M) or THC (20 μ M) in an oocyte expressing rat ANKTM1. **b**, Activation of human ANKTM1 by THC (20 μ M, top row), cannabinol (20 μ M, bottom row) or mustard oil (20 μ M) in transfected HEK293 cells. **c**, Oocytes expressing rat ANKTM1 were perfused with cannabinol (CB, red), THC (green) or mustard oil (yellow) (20 μ M for 1 min each). Voltage ramps (-90 to $+80$ mV) were

applied every 2 s. Averaged currents ($n = 5$ cells) were normalized to maximal mustard-oil-evoked responses at $+80$ mV. **d**, THC dose-response curve for human ANKTM1 in oocytes ($V_h = -60$ mV) recorded in nominally calcium-free ND96. Currents were normalized to maximal responses evoked by THC (40 μ M). Half-maximal activation (EC_{50}) occurred at 12 ± 2 μ M, $n_{Hill} = 2.3 \pm 0.4$ ($n = 3$). THC solubility is poor at concentrations exceeding 40 μ M, thus preventing full saturation of the dose-response curve.

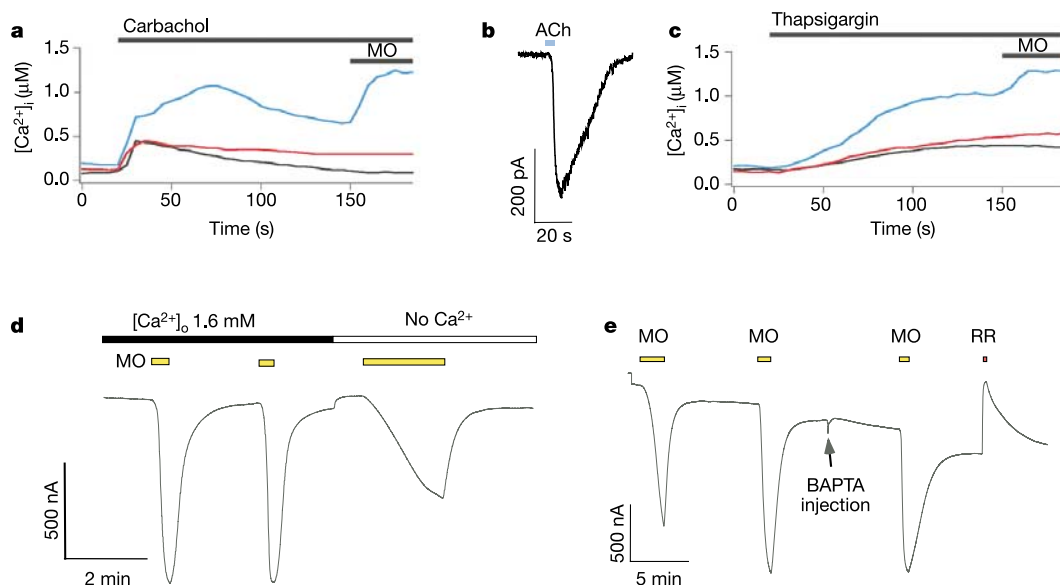


Figure 4 ANKTM1 is a receptor-operated channel. **a**, Application of carbachol (100 μM) to HEK293 cells expressing human ANKTM1 and mAChR evoked large [Ca²⁺]_i responses (blue trace). Co-application of ruthenium red (10 μM, red trace) reduced responses to levels observed in cells transfected with mAChR alone (black trace). ANKTM1 expression was assessed by sensitivity to mustard oil (20 μM). Traces represent averages of three experiments, 35–100 cells in each. **b**, Application of acetylcholine (ACh, 50 μM) produced inward currents in HEK293 cells expressing mAChR1 and human ANKTM1 ($V_h = -60$ mV). **c**, Application of thapsigargin (1 μM) was sufficient to elicit robust increases in [Ca²⁺]_i in ANKTM1-expressing HEK293 cells (blue trace). Co-application of thapsigargin and ruthenium red (10 μM, red trace) reduced responses to the level

observed in vector (pcDNA3)-transfected control (black trace). **d**, Mustard-oil-evoked currents are augmented by extracellular calcium. An ANKTM1-expressing oocyte was activated twice by mustard oil (25 μM) in the presence of extracellular calcium ([Ca²⁺]_o = 1.6 mM), and subsequently in calcium-free medium (0.5 mM EGTA). **e**, Activation of ANKTM1 by mustard oil does not require extracellular or intracellular calcium. Currents evoked by mustard oil (50 μM), measured in calcium-free medium, were identical before and after BAPTA injection (63 nl of a 120 mM solution was sufficient to eliminate receptor-mediated activation of calcium-dependent chloride current; see Supplementary Fig. 3). A slowly developing basal current could be blocked by ruthenium red (5 μM).

intracellular calcium (Fig. 4c). By contrast, cells expressing ANKTM1 showed large and rapid calcium increases that were blocked by ruthenium red (10 μM), suggesting that calcium release is sufficient to activate ANKTM1. We were therefore curious to know whether activation of ANKTM1 by THC or mustard oil is calcium dependent. We found that both the rate and the magnitude of responses to these agonists were enhanced in the presence of extracellular calcium (Fig. 4d). However, THC- or mustard-oil-evoked responses persisted even when both intracellular and extracellular calcium were chelated (Fig. 4e). This shows that calcium is not required for the activation of ANKTM1 by mustard oil or THC, but augments responses to these agonists through an as yet undetermined mechanism. Although these results are consistent with a model in which mustard oil and THC serve as direct channel agonists, they do not exclude an indirect activation mechanism requiring other cellular factors.

We have shown that plants of the genus *Brassica* (for example, wasabi, horseradish and mustard) and *Capsicum* (chilli peppers) use similar molecular strategies to produce irritation and inflammation in herbivorous mammals, both involving activation of excitatory TRP channels on primary sensory nerve endings. Our current findings also extend the concept that some TRP channels may serve as ionotropic cannabinoid receptors, which, in the context of the primary afferent nerve fibre, may contribute to inflammatory hypersensitivity or vasodilation^{5–8,28}. Although THC and cannabinol are natural plant products, they are not known to function as endogenous transmitters in the vertebrate nervous system. In preliminary studies, we have not observed activation of ANKTM1 by known 'endocannabinoids', such as anandamide or *sn*-2-arachidonylglycerol (2-AG), but our findings raise the possibility that other structurally related lipid second messengers may excite nociceptors by targeting this member of the TRP channel family. Finally, our data suggest that ANKTM1

may also serve to excite nociceptors in response to one or more pro-algesic or pro-inflammatory agents—for example, bradykinin, histamine, serotonin, ATP and neurotrophins—that activate PLC signalling pathways in these cells²⁷. Future genetic and pharmacological studies will determine the extent to which this channel contributes to thermal or chemical nociception. □

Methods

Neuronal cell culture, calcium imaging and patch-clamp recording

Trigeminal ganglia from newborn (postnatal day 0) Sprague–Dawley rats were cultured as previously described²⁵. Calcium imaging using Fura-2/AM (Molecular Probes) was analysed using automated routines written in Igor Pro (Wavemetrics). For electrophysiology, gigaseals were formed with pipettes (WPI, Inc.) that had a resistance of 3–5 MΩ in standard pipette solution. Liquid junction potentials (measured in separate experiments) did not exceed 3 mV and thus no correction for this offset was made. Whole-cell voltage clamp was performed at a holding potential of −20 mV with a 200 ms voltage ramp from −100 mV to +80 mV at 3.6 Hz. Data were acquired using Pulse and Pulsefit (HEKA GmbH) software. Recordings were sampled at 20 kHz and filtered at 2 kHz. Extracellular Ringer's solution contained (in mM): 155 NaCl, 4.5 KCl, 2 CaCl₂, 1 MgCl₂, 10 D-glucose and 5 HEPES-Na (pH 7.4). For calcium-free solution, 2 mM MgCl₂ and 1 mM EGTA were substituted for CaCl₂. For mustard oil and muscarinic activation of ANKTM1 in HEK293 cells, recording solutions were as follows (in mM): external, 140 NaCl, 1 MgCl₂, 1 CaCl₂, 1 EGTA and 10 HEPES-Na (pH 7.4); internal, 140 NaCl, 1 MgCl₂, 10 HEPES-Na (pH 7.2), 1 EGTA and 0.2 NaGTP.

ANKTM1 cloning and *in situ* hybridization histochemistry

Rat ANKTM1 sequence was identified by BLAST align of the ENSEMBL and GenBank rat genomic databases with the published mouse sequence (XM_232586.1 and contig RNOR01102533). Full-length complementary DNA (GenBank accession number AY496961) was cloned by polymerase chain reaction (PCR) from an adult rat trigeminal cDNA library using oligonucleotides ATG AAG CGC AGC TTG AGG AGG GTT CTG (5') and CTA GAT GTC TGG GTG GCT AAT AGA ACA ATG TG (3') and KOD polymerase (Novagen–Toboyo). A human ANKTM1 cDNA was kindly provided by B. Trueb³. Apparent mismatches with the ENSEMBL genomic sequence were corrected by PCR and the final product cloned into the pFROG3 mammalian/oocyte expression vector²⁹. *In situ* hybridization histochemistry was performed as previously described¹² using probes corresponding to nucleotides 569–3,378 of the rat ANKTM1 sequence.

Expression in HEK293 cells and oocytes

HEK293 cells were plated on polylysine-coated coverglass or Lab-Tek II CC2 chamberslides (Nalgen-Nunc). Cells were transfected with Lipofectamine 2000 (Invitrogen) using 5–25 ng rat or human ANKTM1 plasmid per cm^2 ; pcDNA3 vector was added to bring the total amount of plasmid DNA to 100 ng cm^{-2} . For muscarinic activation, 100 ng cm^{-2} human mAChR plasmid was co-transfected with 10 ng cm^{-2} ANKTM1 plasmid. Sixteen hours after transfection, cells were loaded with Fura-2AM (5 μM) for 30 min and imaged in Ringer's solution. For oocyte expression, constructs were linearized with *MluI* and transcribed with T7 polymerase (Ambion). Oocytes were incubated in ND96 containing 5 μM ruthenium red as described⁴. Currents were recorded in ND96 (96 mM NaCl, 2 mM KCl, 1.8 mM CaCl_2 , 1 mM MgCl_2 and 5 mM HEPES pH 7.6) or calcium-free ND96 (containing 100 μM BaCl_2) as indicated.

Preparation of natural extracts

For brown mustard, 3 g of ground brown mustard seeds were suspended in 10 ml ND96 and incubated for 2 h at room temperature. Extract was cleared by centrifugation at 3,000g for 20 min. The supernatant fluid was filtered through a 0.2 μm filter and diluted 20-fold in ND96. For wasabi, 0.5 g of 100% pure wasabi paste (Pacific Farms, Oregon) was suspended in 1 ml ND96, clarified by centrifugation at 18,000g, filtered and diluted 50-fold in ND96.

Received 7 November; accepted 12 December 2003; doi:10.1038/nature02282.
Published online 7 January 2004.

1. Wall, P. D. & Melzack, R. *The Textbook of Pain* (W. B. Saunders, London, 1999).
2. Jancso, N., Jancso-Gabor, A. & Szolcsanyi, J. Direct evidence for neurogenic inflammation and its prevention by denervation and by pretreatment with capsaicin. *Br. J. Pharmacol.* **31**, 138–151 (1967).
3. Jaquemar, D., Schenker, T. & Trueb, B. An ankyrin-like protein with transmembrane domains is specifically lost after oncogenic transformation of human fibroblasts. *J. Biol. Chem.* **274**, 7325–7333 (1999).
4. Story, G. M. *et al.* ANKTM1, a TRP-like channel expressed in nociceptive neurons, is activated by cold temperatures. *Cell* **112**, 819–829 (2003).
5. Zygmunt, P. M. *et al.* Vanilloid receptors on sensory nerves mediate the vasodilator action of anandamide. *Nature* **400**, 452–457 (1999).
6. Bisogno, T. *et al.* Molecular targets for cannabidiol and its synthetic analogues: effect on vanilloid VR1 receptors and on the cellular uptake and enzymatic hydrolysis of anandamide. *Br. J. Pharmacol.* **134**, 845–852 (2001).
7. Di Marzo, V., Bisogno, T. & De Petrocellis, L. Anandamide: some like it hot. *Trends Pharmacol. Sci.* **22**, 346–349 (2001).
8. Ross, R. A. Anandamide and vanilloid TRPV1 receptors. *Br. J. Pharmacol.* **140**, 790–801 (2003).
9. Lembeck, F. & Holzer, P. Substance P as neurogenic mediator of antidromic vasodilation and neurogenic plasma extravasation. *Naunyn Schmiedeberg's Arch. Pharmacol.* **310**, 175–183 (1979).
10. Louis, S. M., Jamieson, A., Russell, N. J. & Dockray, G. J. The role of substance P and calcitonin gene-related peptide in neurogenic plasma extravasation and vasodilatation in the rat. *Neuroscience* **32**, 581–586 (1989).
11. Szallasi, A. & Blumberg, P. M. Vanilloid (capsaicin) receptors and mechanisms. *Pharmacol. Rev.* **51**, 159–212 (1999).
12. Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
13. Tominaga, M. *et al.* The cloned capsaicin receptor integrates multiple pain-producing stimuli. *Neuron* **21**, 531–543 (1998).
14. Montell, C., Birnbaumer, L. & Flockerzi, V. The TRP channels, a remarkably functional family. *Cell* **108**, 595–598 (2002).
15. Jordt, S. E., McKemy, D. D. & Julius, D. Lessons from peppers and peppermint: the molecular logic of thermosensation. *Curr. Opin. Neurobiol.* **13**, 487–492 (2003).
16. Clapham, D. E. TRP channels as cellular sensors. *Nature* **426**, 517–524 (2003).
17. Reeh, P. W., Kocher, L. & Jung, S. Does neurogenic inflammation alter the sensitivity of unmyelinated nociceptors in the rat? *Brain Res.* **384**, 42–50 (1986).
18. Caterina, M. J. *et al.* Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science* **288**, 306–313 (2000).
19. Jancso, G., Kiraly, E. & Jancso-Gabor, A. Pharmacologically induced selective degeneration of chemosensitive primary sensory neurones. *Nature* **270**, 741–743 (1977).
20. Simons, C. T., Carstens, M. I. & Carstens, E. Oral irritation by mustard oil: self-desensitization and cross-desensitization with capsaicin. *Chem. Senses* **28**, 459–465 (2003).
21. Caterina, M. J. & Julius, D. The vanilloid receptor: a molecular gateway to the pain pathway. *Annu. Rev. Neurosci.* **24**, 487–517 (2001).
22. Zygmunt, P. M., Andersson, D. A. & Hogestatt, E. D. Δ^9 -Tetrahydrocannabinol and cannabinol activate capsaicin-sensitive sensory nerves via a CB1 and CB2 cannabinoid receptor-independent mechanism. *J. Neurosci.* **22**, 4720–4727 (2002).
23. Wilson, R. K., Kwan, T. K., Kwan, C. Y. & Sorger, G. J. Effects of papaya seed extract and benzyl isothiocyanate on vascular contraction. *Life Sci.* **71**, 497–507 (2002).
24. Patacchini, R., Maggi, C. A. & Meli, A. Capsaicin-like activity of some natural pungent substances on peripheral endings of visceral primary afferents. *Naunyn Schmiedeberg's Arch. Pharmacol.* **342**, 72–77 (1990).
25. McKemy, D. D., Neuhauser, W. M. & Julius, D. Identification of a cold receptor reveals a general role for TRP channels in thermosensation. *Nature* **416**, 52–58 (2002).
26. Peier, A. M. *et al.* A TRP channel that senses cold stimuli and menthol. *Cell* **108**, 705–715 (2002).
27. Rang, H. P., Bevan, S. & Dray, A. in *Textbook of Pain* (eds Wall, P. D. & Melzack, R.) 57–78 (Churchill-Livingstone, London, 1994).

28. Smart, D. *et al.* The endogenous lipid anandamide is a full agonist at the human vanilloid receptor (hVR1). *Br. J. Pharmacol.* **129**, 227–230 (2000).
29. Jordt, S. E., Tominaga, M. & Julius, D. Acid potentiation of the capsaicin receptor determined by a key extracellular site. *Proc. Natl Acad. Sci. USA* **97**, 8134–8139 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to G. Hollopeter for generation of the adult rat trigeminal cDNA library and to B. Trueb for providing us with human ANKTM1 cDNA. This work was supported by grants from the Segerfalk Foundation and the Swedish Research Council (P.Z. and E.H.), the American Heart Association (H.C.) and the National Institutes of Health (I.M., D.B. and D.J.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.J. (julius@cmp.ucsf.edu).

erratum

Structure and conserved RNA binding of the PAZ domain

Kelley S. Yan, Sherry Yan, Amjad Farooq, Arnold Han, Lei Zeng & Ming-Ming Zhou

Nature **426**, 469–474 (2003).

In this Letter, the last two sentences of the second paragraph of the left column on page 470 should read: “Preliminary data using an immobilized 5′ phosphorylated ssRNA with a 3′-biotin modification showed that this RNA failed to pull down purified Ago1 PAZ domain (data not shown). By NMR titration, this RNA had a reduced binding affinity (K_d greater than 20 μM), suggesting that the 3′ end may also play a role in PAZ domain interaction.” □

corrigendum

Eya protein phosphatase activity regulates Six1–Dach–Eya transcriptional effects in mammalian organogenesis

Xue Li, Kenneth A. Ohgi, Jie Zhang, Anna Krones, Kevin T. Bush, Christopher K. Glass, Sanjay K. Nigam, Aneel K. Aggarwal, Richard Maas, David W. Rose & Michael G. Rosenfeld

Nature **426**, 247–254 (2003).

In this Article, K. A. Ohgi's surname was misspelled. It is presented correctly here and has been amended in the HTML version of the paper on *Nature's* website (<http://www.nature.com/nature/>). □

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

Netherlands/ Italy/ Spain/ Portugal/ Belgium:

Evelina Rubio Hakansson (4973)
Scandinavia: Sille Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)
Natureevents:
Paul Constant (4954)

Production Manager:

Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harajukutamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejpn.com

naturejobs

Starting the brain gain

The United Kingdom and Ireland are trying to turn the recruitment tables on the United States — and the US stance on immigration may be inadvertently helping (see page 190). Over the past few years, several programmes, such as Science Foundation Ireland and the Royal Society-Wolfson Research Merit Award, have made inroads in reversing the brain drain by bringing back scientists from abroad. But two new programmes aim to go beyond repatriation.

One such scheme, sponsored by the UK government, looks poised to attract some students who, before the antiterrorist clamp-down, may have considered going to the United States. The Dorothy Hodgkin Postgraduate Awards, announced in November, will fully fund more than 100 graduate students from developing countries to study in Britain. The scheme has a budget of £10 million (US\$18 million) and will target young scientists from India, China, Russia and Hong Kong — some of the countries from which students had the most trouble getting visas into the United States after the terrorist attacks on 11 September 2001. The awards go further than conventional scholarships, as they allow students to stay in Britain for a year after they complete their studies — long enough for them to be recruited by companies or institutions.

And Ireland is hoping to attract scientists in the United States who have biotech management experience or are interested in starting up companies. The government agency Enterprise Ireland last year launched BioNetwork USA-Ireland, which aims to use tax breaks to attract US-based life scientists and managers (see S. Louët *Bioentrepreneur* doi: 10.1038/bioent758).

These programmes challenge other countries to come up with similar schemes to attract the best researchers — and perhaps to make them think about ways to satisfy the ones they already have.

Paul Smaglik
Naturejobs editor



Contents

SPECIAL REPORT

Destressing in the
geekosphere p268

CAREER VIEW

Bricks & Mortar
Lancaster Environment Centre
Graduate Journal
At the crossroads
Movers
Sixto Gonzalez p270

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Finding a balance in the daily grind requires creativity and a sense of play. Kendall Powell explores how lifestyle can complement science.



Life in the geekosphere: with music, food, books, toothbrush and (somewhere) a unicycle, Gordon Kindlmann rarely needs to leave his work space.

Gordon Kindlmann's desk spills onto the floor, invades his neighbour's space and displays the essentials for graduate-student survival — cola cans, leftovers for dinner, antacid bottles, toothbrush, reference books and a unicycle to get around on. Kindlmann, who studies computer science at the University of Utah in Salt Lake City, sits among the chaos with his laptop and his CDs.

"I call this my geekosphere," Kindlmann says fondly. The random oddments on his desk give him something to fiddle with while he ponders his next programming move. But more importantly, the geekosphere creates a sense of self-sufficiency. "In a submarine, I'd be fine for a little while," he says.

It is also a testament to his dedication — so much so, that his adviser brings recruits by to stress the hard work put in by the research group. The long hours, working weekends and occasional tedium of research cut across all disciplines of science and characterize the graduate student and postdoctoral fellow's existence.

The lifestyle makes science less of a nine-to-five job

and more of a zen-like calling. But gruelling hours and repetitive, often frustrating experiments can turn a passion for science into an obsession or an all-out stressfest. Smart lab workers and directors have found creative outlets to improve morale, break up the day, spark scientific ideas and just let off steam. From an early-morning swim to late-night radio, a touch of 'culture' during the working day can inspire and unify labs or provide a brief escape to the outside world.

N. GALLI

ALL IN A DAY'S (AND NIGHT'S) WORK

After the morning's ritual of coffee and e-mail in the University of Toronto's Advanced Center for Detection of Cancer (or AC/DC Lab for short), it's time to get down to some serious work before lunch. Inevitably, someone turns on the lab stereo. In the AC/DC Lab, the tunes are likely to be — you guessed it — heavy metal. The lab's director, clinical biochemist Eleftherios Diamandis, chooses Metallica, the Allman Brothers and Led Zeppelin to drive up the lab's energy level. "Listening increases brain activity," he says, citing his version of the 'Mozart effect' — the idea that musical training may improve memory.

Nicknamed 'Elvis', Diamandis led a rousing game of 'name that tune' at the Gordon conference on cancer detection and diagnosis in Andover, New Hampshire, last August, using the 35,000 digital music files on his laptop, which also help him relax. "It's almost like a live performance. I'm almost in the amphitheatre," he says.

Kindlmann takes a different stance from Diamandis on music's role. He listens to electronic, ambient or post-rock — no words — while he programs. "When you are working, part of your brain wants to be somewhere else," he says. "Music helps to tie that part down and preoccupy it."

By midday, with the gels humming along or the computer crunching data sets, it's time to find the lab adviser to share the week's puzzling results. In Jim Gimzewski's lab at the University of California, Los Angeles, you're likely to find him cross-legged on a brightly carpeted floor with charts spread all around.

Gimzewski, a nanoscientist, specifically designed the PicoLab to look "more like kindergarten" than a typical lab. He furnished rooms with comfy leather and beanbag chairs, carpets, boldly painted walls and striking images of the outdoors to improvise as 'windows' in the basement. He strives for an environment that will provoke new thoughts in lab members as they play with data and chemical models. Gimzewski says that



Students getting plastered? Graduates are immortalized on Erik Jorgensen's lab walls.

nanoscience inspired the "happy future" vision to bring back the "child element of raw creativity". He believes that the comfortable surroundings will help workers feel at home in a lab where they spend most of their waking hours.

Other labs, such as ETH in Zurich, Switzerland, have also forged relaxing havens for their workers (see also *Nature* **424**, 718–720 and 858–859; 2003). Postdocs can retreat to the Raum der Stille, or 'chamber of silence'. The wooden shell lends itself to peaceful meditation — or a prayer to the gods of failed experiments.

By day's end, after the post-lunch dip and caffeine-induced buzz, it is time to regroup for the second shift — the experiments that must continue after most other workers have left for home. While finishing his thesis at the University of Pennsylvania in Philadelphia, Eric Ostertag would hit the gym with colleagues from his molecular-biology lab for a late-afternoon weightlifting workout.

"It was nice to break up the day when we were working 12 or 16 hours," says Ostertag, now a clinical pathology resident at Pennsylvania. Another student shared his flexible training regime, which was ideal for lab work. The getaways provided stress relief and camaraderie, and certainly did not reduce the lab's productivity — Ostertag had published 13 papers by the time he graduated.

EXTRACURRICULAR KICKS

In Erik Jorgensen's lab at the University of Utah, achievements such as Ostertag's would be greeted by the whole group standing on chairs singing a Danish drinking song. The tune (similar to *Oh My Darlin' Clementine* followed by three cheers) celebrates a good result, a publication or just a lab reunion.

Jorgensen, a neurobiologist, instituted another tradition when his first student graduated — at the leaving party, a plaster cast was made of her face. "These people spend six years with you, they are as much family as anything else," he says. Now, a dozen ghostly casts hang in a circle in the lab's lunchroom.

Jorgensen says that although the traditions evolved by accident, they now have a deeper meaning for the lab. The faces are "an exhortation to excel" and "immortalize" successful lab members, he half-jokes.



Lab parties thrown by Martin Giurfa, a behavioural neurobiologist at Paul Sabatier University in Toulouse, France, usually require less plastering and more cooking. He invites his students to share ethnic dishes at a movie night at his home about once a month, and the parties sometimes result in sing-alongs, with Giurfa on the guitar. The lab brings together workers from Asia, Europe and South America, and Giurfa says that the parties help to ease the isolation and culture shock

that many researchers experience while working in a foreign country.

Lab directors say that traditions and celebrations recognize the great commitment of their teams. "We all need reminders that this is somewhat greater than a job, that we are members of that brother- or sisterhood," says Jorgensen.

Sisterhood, for the women in Kai Zinn's molecular-neurobiology laboratory at the California Institute of Technology in Pasadena, includes a lot of conversation. While working late at night 'pushing flies' from one food vial to another, they would watch a small TV or listen to public radio. Discussions might drift over celebrity fashions or ignite into heated political debates, but they

all make time pass more quickly.

Anna Salazar, a graduate student in the lab, says that the cultural conversations serve two other purposes as well. "Sometimes when you are just doing science work, you forget about what's going on outside." Art, music and politics, she says, put her "into creative mode, to get at interesting questions in science".

At the end of a long day, Stephen Voss, a senior lab technician in a cancer-immunology lab at the Mayo Clinic in Rochester, Minnesota, may have discovered the best way to balance mind and body. Voss has trained for 13 years in kendo, the ancient form of Japanese fencing practised by samurais. Kendo uses bamboo sticks and requires discipline beyond the physical sport, including supreme concentration and a code of ethics.

Voss says that it teaches him patience in the face of lab work: "If I have to repeat an experiment, I don't get bent out of shape." Voss — now a kendo master, or *sensei*, teaching his own classes — also competes nationally in choreographed duals fought with real blades and no armour. "If you are doing your kendo correctly, you cannot have external distractions," he says.

Other weekend warriors agree that adventure sports or musical performances leave no mental room for nagging details or worries. Those breaks can provide clarity for a fresh idea to rise to the surface back in the lab. But Zinn says that when he is rock-climbing or hiking, he is still thinking about experiments, forever his favourite pastime.

He and other principal investigators realize that investing in their lab's working environment and promoting balanced lifestyles among the people there pays off in the end. "Science isn't any kind of guaranteed occupation," says Zinn. "People are there because they enjoy doing it. They should control their own experience and decide how they want to do their work." ■

Kendall Powell is a freelance writer based in Broomfield, Colorado.

GRADUATE JOURNAL

At the crossroads

As I approach the halfway point in my PhD, the question about what to do after graduate school is becoming more pressing. My decision is between pursuing my academic research further, or taking an entrepreneurial path. The choice may ultimately depend on the success of my research — altering the catalytic properties of enzymes, an important aspect in biocatalysis. If my experiments succeed, they may open academic doors; if not, there are other options that I would like to explore, such as founding a company.

The entrepreneurial bug comes from friends and colleagues who have started their own businesses. This was particularly endemic before the bubble burst two years ago, but Switzerland still seems to have the right spirit and infrastructure for biotech to thrive. I can envisage setting up a company here with similarly motivated people.

Starting a company requires further skills in addition to the chemistry and biology with aspects of physics that my research involves. So I have not only had to go back to school, but I am also learning to communicate with people from other disciplines who have expertise in different topics to me. I am sure this experience will be of use in the future — regardless of the choice I make when I finish my PhD. ■

Philipp Angerer is a second-year PhD student in biotechnology at the Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.

BRICKS & MORTAR

Lancaster Environment Centre

Last month, what will become one of the largest environmental science centres in Europe opened its doors in Britain. When it's fully staffed, the Lancaster Environment Centre will employ some 300 scientists, including graduate students and postdocs in five departments.

The facility, costing £25 million (US\$46 million), combines five environmental-science departments from Lancaster University with the Centre for Ecology and Hydrology, run by the Natural Environment Research Council (NERC).

The centre is designed to emphasize a multidisciplinary approach to environmental science, says its director, Bill Davies. For example, its five departments can approach different aspects of complex problems, such as

how the genetic make-up of different plants can affect their nitrogen uptake, or how persistent organic pollutants — volatile compounds released by paints — build up in the atmosphere, move around the world and end up in humans' fatty tissue. The project is noteworthy because of the way it combines both government and academic labs, takes a multidisciplinary approach to environmental science, with a strong social-science component, and creates a critical mass of environmental scientists. "We're one of the biggest groups of its kind in Europe," says Davies.

The building, funded jointly by the NERC and Lancaster University, includes controlled environment facilities, 15 glasshouses and equipment to allow integrative studies of terrestrial, aquatic and atmospheric systems.

It is also taking the

kind of computational approach to environmental science that has been the norm in physics and is becoming increasingly common in biology.

Davies says that he is now recruiting a chair in informatics and wants to develop environmental informatics as a specific unit for the centre, with the possibility of adding a new wing to the centre later.

Davies also envisages spinning out some businesses from the site, in much the same way that biotech companies spring from university life-science departments. For instance, scientists at the centre could do contract work on environmental chemistry, or computer modelling of regional environmental problems such as flood prediction. "We can see ways of building on that commercially," explains Davies. ■

► www.lec.lancs.ac.uk

Paul Smaglik is editor of Naturejobs.

MOVERS

Sixto Gonzalez, director, Arecibo Observatory, Puerto Rico



When Sixto Gonzalez left the Massachusetts Institute of Technology (MIT) for his native Puerto Rico midway through his undergraduate education, little did he realize that it would benefit his career.

Gonzalez left, he says, because although he had little difficulty adjusting to a different language, he struggled with the heavy workload in a highly competitive environment — not to mention the long, cold winters, at odds with his native tropical climate.

But a choice based partly on culture shock was served up with a healthy

dose of serendipity. While at the University of Puerto Rico he was able to work at Arecibo Observatory, home of the world's largest and most sensitive single-dish radio telescope — and he has remained there ever since. He says that if he had stayed at MIT, he would never have had the opportunity to work at the observatory.

The opportunities were plentiful. As an undergraduate, he was able to present papers at international conferences based on his findings at Arecibo. As a graduate student, he was able to base his thesis on data he gleaned from a summer spent at the observatory. And as a postgraduate, his familiarity with the instrumentation at Arecibo led him to getting a job there — making him the first native-born Puerto Rican to become a staff scientist at the facility.

His latest appointment extends that honour to the directorship of the

observatory, although Gonzalez has mixed feelings about distinctions based on his nationality. On one hand, he is proud of his heritage and is happy to represent it. But on the other, being singled out by his nationality makes him uncomfortable. He is more excited about having two more Puerto Rican scientists joining him at the facility than about becoming the first director to hail from the country.

Gonzalez says that he is enjoying being in a position to help others to deal with the difficulties he faced early in his career at MIT. Young scientists need a support network, mentoring, ideas for projects and help to bring projects into fruition.

He sees his latest role as helping to provide all these things. "When you're in a position to make a difference for people, you have to provide them with as many opportunities as you can give them," Gonzalez says. ■

CV

2001–2003: Assistant director of Space and Atmospheric Sciences, National Astronomy and Ionosphere Center,

Arecibo, Puerto Rico

1999–2003: Senior research associate, Arecibo Observatory, Puerto Rico

1993–1999: Research associate, Arecibo Observatory